PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-244597
(43)Date of publication of application : 19.09.1995

(51)Int.Cl. G06F 11/20
G06F 3/06
G06F 13/00

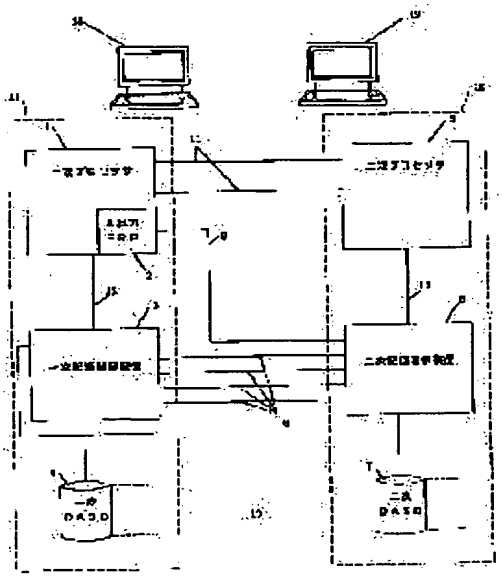
(21)Application number : 06-301146
(22)Date of filing : 05.12.1994

(71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(72)Inventor : KERN ROBERT F
KERN RONALD MAYNARD
MCBRIDE GREGORY EDWARD
SHACKELFORD DAVID MICHAEL

(30)Priority
Priority number : 94 199444 Priority date : 22.02.1994 Priority country : US

(54) METHOD FOR FORMING CONSISTENCY GROUP TO PROVIDE DISASTER RECOVERY FUNCTION, AND ITS RELATED SYSTEM

(57)Abstract:
PURPOSE: To provide a remote data shadowing system which provides a real- time disaster recovery function on a storage area base.
CONSTITUTION: A write input-output operation is performed in a storage sub-system on the primary side 14 by record update on the primary side 14. A time-stamp is attached to this write input-output operation and the time, order and physical position of the record update are collected in a primary data mover. The primary data mover divides plural sets of record update and their related control information into groups based on prescribed time intervals, adds a prefix header to the record update and thereby forms a self-description



record set. The self-description record set is sent to a remote secondary side 15, and such a consistency group is formed that the record update is ordered to be able to shadow the record update in the sequence that matches the sequence where the write input-output operation was performed on the primary side 14 by the record update.

LEGAL STATUS

[Date of request for examination]	05.12.1994
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3149325
[Date of registration]	19.01.2001
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

(19)日本国特許庁（ＪＰ）　(12)特　許　公　報（Ｂ２）　(11)特許番号
特許第3149325号
(P3149325)

(45)発行日　平成13年３月26日(2001.3.26)　(24)登録日　平成13年１月19日(2001.1.19)

(51)Int.Cl.'		識別記号	FI		
Ｇ０６Ｆ	11/20	３１０	Ｇ０６Ｆ	11/20	３１０Ｃ
	3/06	３０４		3/06	３０４Ｆ
	12/00	５３３		12/00	５３３Ｊ
	12/16	３１０		12/16	３１０Ｍ

請求項の数10(全 29 頁)

(21)出願番号　特願平6－301146

(22)出願日　平成６年12月５日(1994.12.5)

(65)公開番号　特開平7－244597

(43)公開日　平成７年９月19日(1995.9.19)

審査請求日　平成６年12月５日(1994.12.5)

(31)優先権主張番号　１９９４４４

(32)優先日　平成６年２月22日(1994.2.22)

(33)優先権主張国　米国（ＵＳ）

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

(72)発明者　ロバート・フレデリック・カーン
アメリカ合衆国85730　アリゾナ州ツーソン　イースト・カレヒコ・ストリート　8338

(74)復代理人　100065455
弁理士　山本　仁朗　（外２名）

審査官　多賀　実

最終頁に続く

(54)【発明の名称】　災害復旧機能を提供するために整合性グループを形成する方法および関連するシステム

(57)【特許請求の範囲】

【請求項１】一次データ・ムーバおよびレコード更新を発生する複数のアプリケーションを実行する一次プロセッサを備える一次側と、一次プロセッサと通信可能に接続された二次プロセッサを備える二次側とを含む災害復旧のための遠隔データ・シャドーイング・システムであって、一次プロセッサが一次記憶サブシステムに連結され、一次プロセッサから一次記憶サブシステムに発する書込み入出力操作に従ってレコード更新を記憶するための記憶装置が一次記憶サブシステムに備えられ、一次側が一次側における時間依存操作を同期させるための共通システム・タイマを更に備えており、二次側が順序整合性のある順序でデータ更新のコピーを記憶するための二次記憶サブシステムを備えているシステムにおける順序整合性のある順序でデータ更新をシャドーイングする方法において、前記方法が、

（ａ）一次記憶サブシステムで発生する各書込み入出力操作にタイム・スタンプを付けるステップと、

（ｂ）一次記憶サブシステムからレコード更新についてのレコード・セット情報を収集するステップと、

（ｃ）データ更新およびレコード・セット情報を一次データ・ムーバに書込んでレコード・セットを生成ステップと、

（ｄ）二次システムにおける書込み入出力操作の順序を再生成するために二次プロセッサにより使用される自己記述レコード・セットを生成するためレコード・セットの各々にヘッダーを付するステップと、

（ｅ）所定の時間間隔に従う時間間隔グループで自己記述レコード・セットを二次プロセッサに転送するステップと、

（ｆ）自己記述レコード・セットの時間間隔グループから整合性グループを形成するステップであって、一次記憶サブシステムに発せられた書込み入出力操作の時間順序に基づいて整合性グループ内においてレコード更新が順序づけされる整合性グループを生成するステップと、
（ｇ）各整合性グループのレコード更新を整合性ある順序で二次記憶サブシステムにシャドーイングするステップと、
（ｈ）ステップ（ｅ）において受取った各自己記述レコード・セットが完全なものであるかどうかを二次側で判定するステップと、
を含む方法。
【請求項２】レコード・セットが二次プロセッサに非同期的に転送されることを特徴とする、請求項１に記載の方法。
【請求項３】ステップ（ｆ）が二次側で行われることを特徴とする、請求項１に記載の方法。
【請求項４】自己記述レコード・セットが不完全であると二次側が判定した場合に、欠落データ更新を再送信するよう二次側が一次側に要求することが、ステップ（ｅ）にさらに含まれることを特徴とする、請求項１に記載の方法。
【請求項５】各時間間隔グループが完全なものであるかどうかを二次側で判定するステップ（ｉ）をさらに含むことを特徴とする、請求項１に記載の方法。
【請求項６】間隔グループが不完全であると二次側が判定した場合に、欠落レコード・セットを再送信するよう二次側が一次側に要求することが、ステップ（ｉ）にさらに含まれることを特徴とする、請求項５に記載の方法。
【請求項７】ステップ（ｂ）が、レコード・セット情報において、各レコード更新が格納されている一次記憶装置上の物理的位置を識別することを特徴とする、請求項１に記載の方法。
【請求項８】ステップ（ｂ）が、レコード・セット情報において、セッション内で一次記憶装置に格納された各レコード更新の順序と更新時間を識別することを特徴とする、請求項７に記載の方法。
【請求項９】ステップ（ｄ）が、接頭部ヘッダにおいて、セッションに関する間隔グループ番号と、そこで参照される各レコード更新のグループ内順序を識別することを特徴とする、請求項１に記載の方法。
【請求項１０】レコード更新を生成するアプリケーションを実行する一次側を含み、一次側から離れた位置に二次側を有し、二次側がレコード更新をシャドーイングして、一次側に災害復旧を提供する、リアルタイム・データ・シャドーイングを行う非同期遠隔データ二重化システムにおいて、非同期遠隔データ二重化システムが、
一次側の時間依存プロセスを同期させるためのシスプレックス・タイマーと、
アプリケーションを実行し、対応するレコード更新用の書込み入出力操作を出し、一次データ・ムーバをそこに有する、一次側の一次プロセッサと、
各レコード更新ごとに書込み入出力操作を１つずつ受け取る複数の一次記憶制御装置であって、それぞれの一次記憶制御装置書込み入出力操作が一次プロセッサによってシスプレックス・タイマーと同期される、複数の一次記憶制御装置と、
対応する書込み入出力操作に応じて、レコード更新をそこに格納するための複数の一次記憶装置とを含み、
一次データ・ムーバが、各レコード更新ごとに複数の一次記憶制御装置からレコード・セット情報を収集して、所定のグループのレコード・セット情報に接頭部ヘッダを付加し、接頭部ヘッダと所定のレコード・セット情報グループが自己記述レコード・セットを形成し、各レコード・セット情報が、一次装置アドレス、シリンダ番号およびヘッド番号（ＣＣＨＨ）、レコード更新順序番号、書込み入出力タイプ、検索引数、セクタ番号、およびレコード更新時間を含み、接頭部ヘッダが、総データ長、操作タイム・スタンプ、時間間隔グループ番号、およびレコード読取り時間を含み、
一次側から自己記述レコード・セットを受け取る二次データ・ムーバを有する二次側の二次プロセッサと、
二次プロセッサに連結された複数の二次記憶制御装置と、
レコード更新を格納する複数の二次記憶装置とをさらに含み、
二次データ・ムーバが、送信された自己記述レコード・セットが完全なものであるかどうかを判定し、自己記述レコード・セットから整合性グループを形成し、複数の一次記憶装置にレコード更新が書き込まれたときの順序に整合する順序で複数の二次記憶装置に書き込むために各整合性グループから得たレコード更新を複数の二次記憶制御装置に出力する、非同期遠隔データ二重化システム。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、一般的には、災害復旧技法に関し、より具体的には、直接アクセス記憶装置（ＤＡＳＤ）データのリアルタイム遠隔コピーのためのシステムに関する。
【０００２】
【従来の技術】データ処理に関連して、効率よくアクセスし、修正および再格納できる大量のデータ（またはレコード）を格納するためには、一般に、データ処理システムが必要である。データ記憶域は、効率よくしかも費用効果の高いデータ記憶を行うために、通常、複数のレベルに、すなわち、階層状態に分かれている。第一のレベル、すなわち、最高レベルのデータ記憶域は、通常はダイナミックまたはスタティック・ランダム・アクセス

・メモリ（ＤＲＡＭまたはＳＲＡＭ）である電子メモリを含む。電子メモリは、ナノ秒単位でこのようなバイト数のデータにアクセスすることによってそれぞれの回路に数百万バイトのデータを格納できる、半導体集積回路の形態をとる。アクセスが完全に電子的に行われるため、このような電子メモリは最高速のデータ・アクセスを提供する。

【０００３】第二レベルのデータ記憶域は、通常、直接アクセス記憶装置（ＤＡＳＤ）を含む。ＤＡＳＤ記憶域は、磁気ディスクまたは光ディスクなどを含む可能性があり、これらのディスクは、データのビットを構成する「１」と「０」を表す磁気的または光学的に変質させたマイクロメートル規模のスポットとしてビット単位のデータをディスク表面に格納する。磁気ＤＡＳＤは、残留磁気材料で被覆した１枚または複数枚のディスクを含む。これらのディスクは、保護環境内に回転式に取り付けられている。各ディスクは、多くの同心トラックまたは間隔が詰まった円に分割されている。データは、各トラックに沿って１ビットずつ順次格納される。ヘッド・ディスク・アセンブリ（ＨＤＡ）として知られているアクセス機構は、一般に、１つまたは複数の読取り書込みヘッドを含み、ディスクが回転して読取り書込みヘッドを通過する際にディスクの表面にデータを転送したりディスクの表面からデータを転送するためにトラック間を移動できるよう、各ＤＡＳＤに設けられている。通常、ミリ秒単位（電子メモリより低速の単位）でこのようなデータにアクセスすることによって、ＤＡＳＤはギガバイト規模のデータを格納できる。所望のデータ記憶位置にディスクおよびＨＤＡを物理的に配置する必要があるため、ＤＡＳＤに格納されたデータへのアクセス速度は低下する。

【０００４】第三またはそれ以下のレベルのデータ記憶域は、テープまたはテープとＤＡＳＤライブラリを含む。このレベルの記憶域では、必要なデータ記憶媒体を選択して装填するのにロボットが必要になるため、ライブラリ内のデータへのアクセス速度はさらに低下する。利点は、テラバイト規模のデータ記憶など、データ記憶容量が非常に大きい割に費用が低減される点である。テープ記憶装置は、バックアップ目的で使用されることが多い。すなわち、階層の第二レベルに格納されたデータは、磁気テープで安全に保管するために複製される。テープまたはライブラリに格納したデータへのアクセス速度は、現在、秒単位である。

【０００５】データの紛失は、企業にとって破滅的なものになる恐れがあるので、多くの企業ではバックアップのデータ・コピーを用意することが必須になっている。一次記憶レベルで紛失したデータの復旧に要する時間も、復旧に関して考慮すべき重要な項目である。テープまたはライブラリでのバックアップ速度の改善策としては、二重コピーが挙げられる。二重コピーの例としては、追加のＤＡＳＤにデータが書き込まれるように追加のＤＡＳＤを用意する方法がある（ミラーリングと呼ばれる場合もある）。この場合、一次ＤＡＳＤで障害が発生しても、データを得るために二次ＤＡＳＤを頼りにすることができる。この方法の欠点は、必要なＤＡＳＤの数が二倍になることである。

【０００６】記憶装置を二重に設ける必要性を克服するもう１つのデータ・バックアップ方法としては、低価格装置の冗長アレイ（ＲＡＩＤ）構成にデータを書き込む方法がある。この方法では、多くのＤＡＳＤにデータを割り当てるようにデータが書き込まれる。１つのＤＡＳＤで障害が発生しても、残りのデータとエラー修正手順を使用すれば、紛失したデータを復旧できる。現在では、数種類のＲＡＩＤ構成が使用できる。

【０００７】一般に、記憶装置または記憶媒体で障害が発生した場合のデータの復旧には、上記のバックアップ方法で十分である。このようなバックアップ方法では、二次データは一次データのミラーになる、つまり、二次データは一次データと同じボリューム通し番号（ＶＯＬＳＥＲ）とＤＡＳＤアドレスを持つので、これらの方法は装置障害の場合だけ有用である。しかし、ミラーリングされた二次データを使用しても、システム障害の復旧は行えない。このため、地震、火災、爆発、台風など、システム全体またはシステム使用現場を破壊する災害が発生した場合にデータを復旧するには、さらに保護が必要である。つまり、災害復旧のためには、一次データから離れた位置にデータの二次コピーを格納しておく必要がある。災害保護を行う既知の方法としては、毎日または毎週、テープにデータをバックアップする方法がある。この場合、車両でテープを収集し、通常、一次データ位置から数キロメートル離れた安全な保管区域にテープを移送する。このバックアップ計画には、バックアップ・データを検索するのに何日もかかると同時に、数時間分または数日分のデータが失われ、最悪の場合、同じ災害によって保管場所も破壊されてしまう恐れがあるという問題がある。多少改善されたバックアップ方法としては、毎晩、バックアップ場所にデータを転送する方法が考えられる。この方法では、より離れた遠隔地にデータを格納することができる。しかし、この場合も、二重コピー方法と同様、バックアップが連続して行われるわけではないので、次のバックアップまでにデータの一部が失われる可能性がある。このため、一部のユーザにとっては受け入れられないほど、相当なデータ量が失われる恐れがある。

【０００８】さらに最近導入されたデータ災害復旧方法としては、遠隔方式だけでなく、連続方式でもデータがバックアップされる、遠隔二重コピーがある。あるホスト・プロセッサから別のホスト・プロセッサへ、またはある記憶制御装置から別の記憶制御装置へ、あるいはこれらを組み合わせた方法で二重データを送信するには、

このプロセスを実現するために相当な量の制御データが必要になる。しかし、オーバヘッドが高いために、二次側が一次側の処理に遅れないようにする能力が妨げられる可能性があり、このため、災害が発生した場合に二次側が一次側を復旧する能力が脅かされる。

【０００９】したがって、最小限の制御データを使用して、一次処理側のデータと一致するデータのリアルタイム更新を行う方法および装置を提供することが望まれている。この方法および装置は、復旧される特定のアプリケーション・データから独立して、つまり、特定のアプリケーション・データをベースとするのではなく、汎用記憶媒体をベースとして機能する。

【００１０】

【発明が解決しようとする課題】本発明の目的は、災害復旧のためにＤＡＳＤデータを二次側にシャドーイング（shadowing）するための改良された設計および方法を提供することにある。

【００１１】

【課題を解決するための手段】本発明の第一の実施例によれば、整合性グループを形成するための方法により、遠隔地からの災害復旧機能が提供される。一次プロセッサで実行される１つまたは複数のアプリケーションによって生成されるデータ更新は、一次記憶サブシステムによって受け取られ、この一次記憶サブシステムは、入出力書込み操作により各データ更新がそこに書き込まれるようにする。一次記憶サブシステムは共通タイマによって同期が取られ、一次プロセッサから離れた位置にある二次システムは、災害復旧のために二次側が使用できるように順序整合性のある順序でデータ更新をシャドーイングする。本方法は、（ａ）一次記憶サブシステムで行われる各書込み入出力操作にタイム・スタンプを付けるステップと、（ｂ）各データ更新ごとに一次記憶サブシステムから書込み入出力操作のレコード・セット情報を収集するステップと、（ｃ）自己記述レコード・セットが一連の書込み入出力操作を再生成するのに十分なものになるように、データ更新とそれぞれのレコード・セット情報から自己記述レコード・セットを生成するステップと、（ｄ）所定の間隔しきい値に基づいて、自己記述レコード・セットを間隔グループにグループ分けするステップと、（ｅ）最も早い操作タイム・スタンプを持つ自己記述レコード・セットの間隔グループとして、第一の整合性グループを選択するステップとを含み、個々のデータ更新が、一次記憶サブシステム内の入出力書込み操作の時間順に基づいて、第一の整合性グループ内で順序付けされる。

【００１２】本発明の他の実施例では、一次システムは１つまたは複数のアプリケーションを実行する一次プロセッサを有し、このアプリケーションがレコード更新を生成し、一次プロセッサがそれに基づく自己記述レコード・セットを生成する。それぞれの自己記述レコード・セットは、一次システムから離れた位置にある二次システムに送られ、二次システムは、リアルタイム災害復旧のために、自己記述レコード・セットに基づいて、順序整合性のある順序でレコード更新をシャドーイングする。一次プロセッサは一次記憶サブシステムに連結され、一次記憶サブシステムはレコード更新を受け取って、入出力書込み操作によって各レコード更新がそこに格納されるようにする。一次プロセッサは、同期を取るためにアプリケーションと一次記憶サブシステムに共通の時間源を提供するためのシスプレックス・クロックを含み、シスプレックス・クロックによって同期が取られる一次データ・ムーバは、各レコード更新ごとに一次データ・ムーバにレコード・セット情報を提供するよう、一次記憶サブシステムに指示する。一次データ・ムーバは、複数のレコード更新とそれに対応する各レコード・セット情報を時間間隔グループにグループ分けし、それに接頭部ヘッダを挿入する。それぞれの時間間隔グループは自己記述レコード・セットを形成する。

【００１３】本発明の上記およびその他の目的、特徴、および利点は、添付図面に図示する本発明の実施例に関する以下の詳細な説明から明らかになるだろう。

【００１４】

【実施例】一般的なデータ処理システムは、データを計算して操作し、データ機能記憶管理サブシステム／多重仮想記憶システム（ＤＦＳＭＳ／ＭＶＳ）ソフトウェアなどを実行するために、ＩＢＭシステム／３６０またはＩＢＭシステム／３７０プロセッサなどのホスト・プロセッサの形態をとり、少なくとも１台のＩＢＭ　３９９０記憶制御装置がそれに接続され、その記憶制御装置が、メモリ制御装置と、それに組み込まれた１つまたは複数のタイプのキャッシュ・メモリとを含む場合がある。さらに記憶制御装置は、ＩＢＭ　３３８０または３３９０　ＤＡＳＤなどの１群の直接アクセス記憶装置（ＤＡＳＤ）に接続されている。ホスト・プロセッサが実質的な計算能力を提供するのに対し、記憶制御装置は、大規模データベースを効率よく転送し、ステージ（stage）／デステージ（destage）し、変換し、全般的にアクセスするのに必要な諸機能を提供する。

【００１５】一般的なデータ処理システムの災害復旧保護では、一次ＤＡＳＤに格納した一次データを二次側または遠隔地でバックアップする必要がある。一次側と二次側との距離は、ユーザが受け入れられる危険のレベルによって決まり、数キロメートルから数千キロメートルの範囲が可能である。二次側または遠隔地は、バックアップ・データ・コピーを提供するだけでなく、一次システムが使用不能になった場合に一次システムの処理を引き継ぐのに十分なシステム情報を持っていなければならない。その理由は、主に、単一の記憶制御装置では一次側と二次側に設けた一次および二次ＤＡＳＤストリングの両方にデータを書き込めないためである。むしろ、一

次記憶制御装置に接続された一次ＤＡＳＤストリングには一次データが格納されるのに対し、二次記憶制御装置に接続された二次ＤＡＳＤストリングには二次データが格納されるのである。

【００１６】二次側は、一次側から十分離れている必要があるだけでなく、一次データをリアルタイムでバックアップできなくてはならない。二次側は、一次データが更新されたときに、最小限の遅延で一次データをバックアップする必要がある。しかも、二次側は、一次側で実行され、データまたは更新を生成するアプリケーション・プログラム（たとえば、ＩＭＳやＤＢ２）を考慮せずに、一次データをバックアップしなければならない。二次側に要求される難しい課題は、二次データの順序が整合していなければならないことである。つまり、二次データは一次データと同じ順序でコピーされなければならず（順序整合性）、これはシステムについてかなり考慮を要する問題である。順序整合性は、それぞれが１つのデータ処理システム内の複数のＤＡＳＤを制御する記憶制御装置が複数存在するためにさらに複雑になっている。順序整合性がないと、一次データと一致しない二次データが生成され、その結果、災害復旧が崩壊する恐れがある。

【００１７】遠隔データの二重化は、同期と非同期の２つの一般的なカテゴリに分けられる。同期遠隔コピーでは、一次データを二次側に送り、一次ＤＡＳＤの入出力操作を終了する（一次ホストにチャネル終了（ＣＥ）と装置終了（ＤＥ）を出力する）前にこのようなデータの受取りを確認する必要がある。このため、同期コピーでは、二次側の確認を待っている間に一次ＤＡＳＤの入出力応答時間が遅くなる。一次側の入出力応答遅延は、一次システムと二次システムとの距離に比例して長くなる（これは遠隔距離を数十キロメートル規模に制限する要素である）。しかし、同期コピーは、システム・オーバヘッドを比較的小さくして、順序が整合したデータを二次側に提供する。

【００１８】非同期遠隔コピーでは、二次側でデータが確認される前に一次ＤＡＳＤの入出力操作が完了する（一次ホストにチャネル終了（ＣＥ）と装置終了（ＤＥ）を出力する）ため、一次側のアプリケーション・システムのパフォーマンスが向上する。このため、一次ＤＡＳＤの入出力応答時間は二次側までの距離に依存せず、一次側から数千キロメートル離れた遠隔地に二次側を設けることもできる。しかし、二次側で受け取ったデータの順序が一次側の更新順序と一致しなくなる場合が多いので、データの順序整合性を確保するのに必要なシステム・オーバヘッドが増加する。したがって、一次側で障害が発生すると、一次側と二次側との間で転送中のデータが一部紛失する恐れがある。

【００１９】同期データ・シャドーイング

災害復旧のための同期リアルタイム遠隔コピーでは、コピーされた複数のＤＡＳＤボリュームが１つのセットを形成する必要がある。さらに、このようなセットを形成するには、各セットを構成するこれらのボリューム（ＶＯＬＳＥＲ）とそれに対応する一次側の同等物を識別するために十分な量のシステム情報を二次側に提供する必要がある。重要なのは、二次側が一次側と「二重対（duplex pair）」を形成するので、１つまたは複数のボリュームがこのセットと同期していない、つまり、「二重対の障害」が発生したときに二次側がそれを認識しなければならない点である。代替経路が再試行される間に一次ＤＡＳＤの入出力が遅延するため、非同期遠隔コピーより同期遠隔コピーの方が、接続障害を認識しやすい。一次側は、二次側用の更新を待ち行列に入れる間、一次側が続行できるようにコピーを中止または中断することができ、一次側は、二次側が同期していないことを示すためにこのような更新にマークを付ける。二次側がいつでも災害復旧に使用できるようにするため、二次側が一次側と同期しなくなる原因になりそうな例外条件を認識することが必要である。エラー条件や復旧処置によって、二次側が一次側と一致しなくなってはならない。

【００２０】しかし、二次ＤＡＳＤが存在しアクセス可能である状態で、二次側と一次側の接続を維持しても、内容の同期は確保されない。いくつかの理由から、二次側は一次側との同期性を失う場合もある。二重対が形成されたときに二次側は当初同期しておらず、初期データ・コピーが完了したときに同期に達する。一次側が二次側に更新済みデータを書き込めない場合、一次側は二重対を解除することもある。この場合、一次側は、更新アプリケーションが続行できるように、二重対が中断された状況で一次ＤＡＳＤに更新内容を書き込む。このため、二重対が復元されるまで、一次側は露出状態で、つまり、現行の災害保護コピーを使わずに実行を続ける。二重対が復元されても、二次側は直ちに同期状態になるわけではない。この時点で保留状態の更新を適用した後で、二次側は同期状態に戻る。一次側は、該当ボリュームに関する中止コマンドを一次ＤＡＳＤに出すことによって、二次側の同期を喪失させることもできる。この中止コマンドが終了し、二重対が再確立され、保留状態の更新がコピーされると、二次側は一次側と再同期する。また、オンライン保守でも、同期を喪失させることができる。

【００２１】二次ボリュームが一次ボリュームと同期していない場合、二次ボリュームは、二次システムの復旧と一次側アプリケーションの再開に使用することができない。二次側の非同期ボリュームは非同期ボリュームとして識別されなければならず、二次側の復旧引継ぎ手順は、アプリケーション・アクセスを否定する（そのボリュームを強制的にオフラインにするか、そのＶＯＬＳＥＲを変更する）ために非同期ボリュームを識別する必要がある。一次側ホストがアクセス不能になった場合に一

次側を復旧するために、二次側が呼び出されることがある。このため、二次側では、すべてのボリュームの同期状態に関するすべての関連情報が必要である。二次記憶サブシステム、すなわち、二次記憶制御装置とＤＡＳＤは、一次側で検出された例外によって一次側が同期を解除する原因となるすべての条件を判別できるわけではない。たとえば、二次側が把握していない一次入出力経路またはリンクの障害のために一次側が二次側の対等機能にアクセスできない場合、一次側は二重対を解除することがある。この場合、二次側が同期状態を示すのに対し、一次側は、二重対が解除されたことを示す。

【００２２】非同期二重対ボリュームが存在することを、外部通信によって二次側に通知することができる。これは、ユーザ・システム管理機能を使用することで認識できる。一次側の入出力操作はチャネル終了／装置終了／装置チェック（ＣＥ／ＤＥ／ＵＣ）状況で終了し、センス・データがエラーの特徴を示す。このような形式の入出力構成の場合、エラー回復プログラム（ＥＲＰ）がエラーを処理し、入出力の完了を一次側アプリケーションに通知する前に二次プロセッサに適切なメッセージを送る。この場合、ＥＲＰの二重対中止メッセージを認識し、その情報を二次側で確保するのは、ユーザの責任である。一次側の代わりに動作可能になるよう二次側が頼りにされている場合は、始動手順によって二次ＤＡＳＤが二次ホストにオンライン接続され、アプリケーションの割振りのために非同期ボリュームがオンライン接続されていないことを確認するために、二次ＤＡＳＤサブシステムに格納された同期状況が検索される。この同期状況をすべてのＥＲＰ二重対中止メッセージと組み合わせると、二次側の非同期ボリューム全体を示すピクチャが得られる。

【００２３】ここで図１を参照して説明すると、同図には、一次側１４と二次側１５を有し、二次側１５が一次側１４から２０キロメートル離れている、災害復旧システム１０が示されている。一次側１４は、そこで実行されているアプリケーションと、システム入出力およびエラー回復プログラム２（以下、入出力ＥＲＰ２という）とを有するホスト・プロセッサまたは一次プロセッサ１を含む。一次プロセッサ１は、ＤＦＳＭＳ／ＭＶＳオペレーティング・ソフトウェアを実行するＩＢＭエンタープライズ・システム／９０００（ＥＳ／９０００）などでもよく、さらにそこで複数のアプリケーション・プログラムを実行することもできる。一次プロセッサ１には、ＩＢＭ　３９９０－６型記憶制御装置などの一次記憶制御装置３がチャネル１２を介して接続されている。当技術分野で既知の通り、複数のこのような一次記憶制御装置３を１台の一次プロセッサ１に接続するか、あるいは複数の一次プロセッサ１を複数の一次記憶制御装置３に接続することができる。一次記憶制御装置３には、ＩＢＭ　３３９０　ＤＡＳＤなどの一次ＤＡＳＤ４が接続されている。複数の一次ＤＡＳＤ４を一次記憶制御装置３に接続することができる。一次記憶制御装置３と、これに接続された一次ＤＡＳＤ４によって、一次記憶サブシステムが形成される。また、一次記憶制御装置３と一次ＤＡＳＤ４は、単一の一体型ユニットであってもよい。

【００２４】二次側１５は、チャネル１３を介してＩＢＭ　３９９０－３型などの二次記憶制御装置６に接続されたＩＢＭ　ＥＳ／９０００などの二次プロセッサ５を含む。二次記憶制御装置６には、さらにＤＡＳＤ７が接続されている。一次プロセッサ１は、チャネル・リンクまたはＴ１／Ｔ３電話回線リンクなどの少なくとも１つのホスト間通信リンク１１によって二次プロセッサ５に接続されている。一次プロセッサ１は、複数のエンタープライズ・システム接続（ＥＳＣＯＮ）リンク９などによって二次記憶制御装置６との直接接続を確保することもできる。その結果、入出力ＥＲＰ２は、必要であれば、二次記憶制御装置６と通信可能になる。一次記憶制御装置３は、複数のＥＳＣＯＮリンクなどの複数の対等通信リンク８を介して二次記憶制御装置６と通信する。

【００２５】一次プロセッサ１で実行されるアプリケーション・プログラムによって書込み入出力操作が実行されると、入出力操作が正常に完了したことを示すハードウェア状況としてチャネル終了／装置終了（ＣＥ／ＤＥ）が出力される。入出力操作が正常に完了すると、一次プロセッサ１のオペレーティング・システム・ソフトウェアは、そのアプリケーションに書込み入出力成功のマークを付け、それにより、アプリケーション・プログラムは、最初または前の書込み入出力操作の正常終了に依存する可能性のある次の書込み入出力操作に移行できるようになる。これに対して、書込み入出力操作が不成功に終わった場合は、チャネル終了／装置終了／装置チェック（以下、ＣＥ／ＤＥ／ＵＣという）という入出力状況が一次プロセッサ１のオペレーティング・システム・ソフトウェアに出力される。装置チェックを出力した後、入出力ＥＲＰ２は、制御権を引き継ぎ、失敗した書込み入出力操作の特徴に関する具体的なセンス情報を一次記憶制御装置３から入手する。あるボリュームに固有のエラーが発生した場合は、そのエラーに関連する固有の状況が入出力ＥＲＰ２に出力される。その後、入出力ＥＲＰ２は、一次記憶制御装置３と二次記憶制御装置６との間、または最悪の場合は、一次プロセッサ１と二次プロセッサ５との間のデータ保全性を維持するために、新たな対等通信同期エラー回復を実行することができる。

【００２６】図２および図３を参照して説明すると、同図にはエラー回復手順が示されている。図２のステップ２０１は、一次プロセッサ１で実行されるアプリケーション・プログラムが一次記憶制御装置３にデータ更新を送信することを含む。ステップ２０３では、そのデータ

更新が一次ＤＡＳＤ４に書き込まれ、そのデータ更新が二次記憶制御装置６にシャドーイングされる。ステップ２０５では、二重対の状況がチェックされ、一次側と二次側が同期しているかどうかが判別される。二重対の状況が同期状態になっている場合、ステップ２０７でデータ更新が二次ＤＡＳＤ７に書き込まれ、一次プロセッサ１での処理は、そこで実行されるアプリケーション・プログラムを介して続行される。

【００２７】二重対が「障害発生」状態になっている場合、ステップ２０９で一次記憶制御装置３は、二重対で中断または障害が発生していることを一次プロセッサ１に通知する。二重対は、通信リンク８による一次記憶制御装置３と二次記憶制御装置６との通信障害によって「障害発生」状態になる場合がある。あるいは、二重対は、一次サブシステムまたは二次サブシステムいずれかのエラーによって「障害発生」状態になる場合もある。障害が通信リンク８で発生している場合、一次記憶制御装置３は、二次記憶制御装置６に直接、障害を連絡することができない。そこで、一次記憶制御装置３は、ステップ２１１で入出力状況としてＣＥ／ＤＥ／ＵＣを一次プロセッサ１に返す。入出力ＥＲＰ２は、アプリケーション・プログラムを静止させ、書込み入出力操作を要求するアプリケーションに制御権を返す前に、エラー回復とデータ保全性のためにステップ２１３で一次プロセッサ１の制御権を引き継ぐ。

【００２８】図３は、入出力ＥＲＰ２が実行する諸ステップを表している。ステップ２２１で入出力ＥＲＰ２は一次記憶制御装置３にセンス入出力を出す。センス入出力操作は、入出力エラーの原因を記述する情報を返す。すなわち、このデータ記述情報は、具体的なエラーに関して記憶制御装置または二重対の操作に固有のものになる。一次記憶制御装置３と二次記憶制御装置６との間の対等通信リンク８で障害が発生したことがデータ記述情報によって示された場合、ステップ２２３で入出力ＥＲＰ２は、一次記憶制御装置３および二次記憶制御装置６に対して、関係ボリュームを同期遠隔コピー障害状態に入れるように指示する記憶制御装置レベル入出力操作を出す。この二次記憶制御装置６は、複数のＥＳＣＯＮリンク９またはホスト間通信リンク１１を介して入出力ＥＲＰ２から関係ボリュームの状態を受け取ることができる。その結果、二重対操作の現在の状況は、一次プロセッサ１で実行されるアプリケーションとともに、一次プロセッサ１および二次プロセッサ５の両方で維持される。コンソール１８および１９は、それぞれ一次プロセッサ１および二次プロセッサ５からの情報をやりとりするために設けられ、入出力ＥＲＰは、両方のコンソール１８および１９に状況情報を通知する。

【００２９】一次記憶制御装置３および二次記憶制御装置６への同期遠隔コピー入出力操作障害が正常に完了したとき、ステップ２２５ではデータ保全性が維持されている。このため、二次側１５で復旧を試みると、二次記憶制御装置６は、「同期遠隔コピー障害」というマークを付けたボリュームを、データ回復手段（ボリューム上のそのデータの状態を判別するための従来のデータベース・ログまたはジャーナル）によってそのボリュームのデータとその同期グループ内の他のデータとの同期が取られるまで使用できないものとして識別する。

【００３０】ステップ２２７では、同期遠隔コピー障害の状況更新について一次記憶制御装置３と二次記憶制御装置６で行われた入出力操作の正常終了を入出力ＥＲＰ２が受け取ったかどうかを判別するテストが行われる。正常終了すると、入出力ＥＲＰ２は、ステップ２２９で一次プロセッサ１に制御権を返す。正常終了していない場合は、ステップ２３１で次のレベルの復旧通知が行われる。この通知には、障害発生ボリュームと、一次記憶制御装置３または二次記憶制御装置６のいずれかのそのボリュームの状況が正しくない可能性があることを、コンソール１８を介してオペレータに通知することが含まれる。この通知は、そこで具体的なボリューム状況を示すために、コンソール１９または共用ＤＡＳＤデータ・セットを介して二次側１５にシャドーイングされる。

【００３１】ステップ２３３で、エラー・ログ記録データ・セットが更新される。この更新は、一次ＤＡＳＤ４または他の記憶場所のいずれかに書き込まれ、二次側１５にシャドーイングされる。このエラー回復処置が完了すると、入出力ＥＲＰ２はステップ２３５で、書込み入出力操作障害に関する「永続エラー」回復を一次側アプリケーションに実行させるために、一次側アプリケーションの書込み入出力操作に「永続エラー」を通知する。エラーが修正されると、ボリューム状態は、まず保留状態（変更データの再コピー）に回復し、次に全二重に回復することができる。その後、二重対が再確立されると、データを二次ＤＡＳＤ７に再適用することができる。

【００３２】二重対を確立する場合、顧客の要求に応じて、ボリュームをＣＲＩＴＩＣＡＬと識別することができる。ＣＲＩＴＩＣＡＬボリュームの場合、ある操作の結果、二重対の障害が発生すると、実際のエラー箇所とは無関係に、一次ボリュームの永続エラー障害が報告される。ＣＲＩＴ＝Ｙの場合、障害発生対の一次ＤＡＳＤ４０６に書き込もうとするその後のすべての試みは、永続エラーを受け取ることになり、対をなす二次ボリュームにシャドーイングできないデータは、その一次ボリュームに一切書き込まれなくなる。このため、必要であれば、一次側アプリケーションの処置および入出力データ操作との完全同期が可能になる。

【００３３】その結果、本明細書に記載する災害復旧システム１０では、入出力命令（チャネル・コマンド・ワード（ＣＣＷ））を有する一次ホスト処理エラー回復手順によって、一次および二次同期遠隔コピー・ボリュー

ムの状況を二重対から障害発生二重対へ変更できるようにする同期遠隔コピーを取り入れ、それにより、複数タイプの一次および二次サブシステム・エラーの場合にデータ保全性を維持する。アプリケーション・ベースのバックアップではなく、データ更新がリアルタイムで複写される記憶域ベースのバックアップが設けられている。また、災害復旧システム１０は、（１）一次および二次記憶制御装置ボリューム状況更新、（２）オペレータ・メッセージまたはエラー・ログ記録共通データ・セットを介して具体的なボリューム更新状況に関して一次および二次ホスト・プロセッサが通知すること、および（３）ＣＲＩＴＩＣＡＬボリューム識別などの、複数レベルの一次／二次状況更新を試み、ボリューム対が障害発生二重対になる場合は、一次ボリュームへのその後の更新を防止することができる。このため、リアルタイムの完全エラー災害復旧が達成される。

【００３４】非同期データ・シャドーイング

非同期遠隔データ・シャドーイングは、１回の災害で一次側と二次側の両方が崩壊してしまう確率を低減するために一次側と二次側との距離をさらに大きくする必要がある場合、または一次側アプリケーションのパフォーマンスへの影響を最小限に抑える必要がある場合に使用する。一次側と二次側との距離は、現在では地球全体またはそれ以上に延長できるが、複数の一次サブシステムの背後にある複数のＤＡＳＤボリュームにわたる書込み更新を複数の二次サブシステムに同期させることは、さらに複雑である。二次記憶サブシステム上でシャドーイングするために、一次データ・ムーバを介して一次記憶制御装置から二次データ・ムーバへレコード書込み更新を発送することができるが、両者間でやりとりされる制御データの量は、最小限でなければならず、同時に、複数の記憶制御装置に隠れている複数のＤＡＳＤボリュームにわたる一次システムの場合と同様、複数の記憶制御装置にわたる二次システム上でレコード書込み更新の順序を正確に再構築できるものでなければならない。

【００３５】図４は、一次側４２１と遠隔側または二次側４３１とを含む非同期災害復旧システム４００を示している。一次側４２１は、ＤＦＳＭＳ／ＭＶＳホスト・ソフトウェアを実行するＩＢＭ　ＥＳ／９０００などの一次プロセッサ４０１を含む。一次プロセッサ４０１は、ＩＭＳおよびＤＢＳアプリケーションなどのアプリケーション・プログラム４０２および４０３と、一次データ・ムーバ（ＰＤＭ）４０４をさらに含む。一次プロセッサ４０１には、そこで実行されるすべてのアプリケーション（４０２、４０３）に共通の基準を提供するために、共通シスプレックス・クロック（sysplex clock）４０７が設けられ、すべてのシステム・クロックまたは時間源（図示せず）がシスプレックス・クロック４０７に同期し、すべての時間依存プロセスが相互に正しいタイミングで動作するようになっている。たとえば、一次記憶制御装置４０５は、単一の一次記憶制御装置４０６への２回の連続する書込み入出力操作が同じタイム・スタンプ値を示さないように、複数のレコード書込み更新時間を確実に区別するのに適した解像度に同期している。シスプレックス・クロック４０７の解像度（正確さではない）は重要である。ＰＤＭ４０４は、シスプレックス・クロック４０７に接続された状態で図示されているが、書込み入出力操作がそこで発生するわけではないので、シスプレックス・クロック４０７に同期させる必要はない。また、一次プロセッサ４０１が単一の時間基準（たとえば、単一のマルチプロセッサＥＳ／９０００システム）を有する場合には、シスプレックス・クロック４０７は不要である。

【００３６】一次プロセッサ４０１には、ＩＢＭ　３９９０－６型記憶制御装置などの複数の一次記憶制御装置４０５が光ファイバ・チャネルなどの複数のチャネルを介して接続されている。また、各一次記憶制御装置４０５には、ＩＢＭ　３３９０ＤＡＳＤなどの複数の一次ＤＡＳＤ４０６からなる少なくとも１つのストリングが接続されている。一次記憶制御装置４０５と一次ＤＡＳＤ４０６によって、一次記憶サブシステムが形成される。各記憶制御装置４０５と一次ＤＡＳＤ４０６は、個別のユニットである必要はなく、両者を組み合わせて単一のドロワにしてもよい。

【００３７】一次側４２１から数千キロメートル離れた位置に配置される二次側４３１は、一次側４２１と同様に、そこで動作する二次データ・ムーバ（ＳＤＭ）４１４を有する二次プロセッサ４１１を含む。あるいは、一次側と二次側が同じ場所に存在してもよく、さらに、一次データ・ムーバと二次データ・ムーバが単一のホスト・プロセッサに常駐してもよい（二次ＤＡＳＤは防火壁のすぐ上に設けてもよい）。二次プロセッサ４１１には、当技術分野で既知の通り、光ファイバ・チャネルなどのチャネルを介して複数の二次記憶制御装置４１５が接続されている。記憶制御装置４１５には、複数の二次ＤＡＳＤ４１６と１つの制御情報ＤＡＳＤ４１７（複数も可）が接続されている。記憶制御装置４１５とＤＡＳＤ４１６および４１７によって、二次記憶サブシステムが構成される。

【００３８】一次側４２１は、通信リンク４０８を介して二次側４３１と通信する。より具体的には、一次プロセッサ４０１は、仮想記憶通信アクセス方式（ＶＴＡＭ）通信リンク４０８などの通信プロトコルによって、二次プロセッサ４１１にデータと制御情報を転送する。この通信リンク４０８は、電話（Ｔ１、Ｔ３回線）、無線、無線／電話、マイクロ波、衛星などの複数の適当な通信方式によって実現できる。

【００３９】非同期データ・シャドーイング・システム４００は、一次ＤＡＳＤ４０６へのすべてのデータ書込みの順序が保持され、二次ＤＡＳＤ４１６に適用される

（すべての一次記憶サブシステムにわたるデータ書込み順序を保持する）ように、一次記憶制御装置４０５から制御データを収集する機能を含む。二次側４３１に送られるデータおよび制御情報は、データ保全性を保持するのに一次側４２１の存在が不要になるほど、十分なものでなければならない。

【００４０】アプリケーション４０２、４０３は、データまたはレコード更新を生成するが、このレコード更新は、一次記憶制御装置４０５によって収集され、ＰＤＭ４０４によって読み取られる。それぞれの一次記憶制御装置４０５は、非同期遠隔データ・シャドーイング・セッションのためにそれぞれのレコード更新をグループ化し、非特定一次ＤＡＳＤ４０６のＲＥＡＤ要求を介してＰＤＭ４０４にこれらのレコード更新を提供する。一次記憶制御装置４０５からＰＤＭ４０４へのレコード更新の転送は、ＳＴＡＲＴ入出力操作の回数および読取りから読取りの遅延を最小限にしながら、各一次記憶制御装置４０５と一次プロセッサ４０１との間で転送されるデータの量を最大にするように、ＰＤＭ４０４によって制御され、最適化される。ＰＤＭ４０４は、非特定ＲＥＡＤ間の時間間隔を変えることで、一次記憶制御装置とホストとのこの最適化だけでなく、二次ＤＡＳＤ４１６用のレコード更新の通用期間も制御することができる。

【００４１】データ保全性を維持しながら、ＰＤＭ４０４がレコード更新を収集し、そのレコード更新をＳＤＭ４１４に送信するには、すべての一次記憶サブシステムにおいて二次ＤＡＳＤ４１６に対して行われる一次ＤＡＳＤ４０６のレコードＷＲＩＴＥシーケンスを再構築するのに十分な制御データとともに、特定の期間の間、適切な複数の時間間隔でレコード更新を送信する必要がある。一次ＤＡＳＤ４０６のレコードＷＲＩＴＥシーケンスの再構築は、自己記述レコードをＰＤＭ４０４からＳＤＭ４１４に渡すことによって達成される。ＳＤＭ４１４は、所与の時間間隔分のレコードが紛失しているかどうか、または不完全になっているかどうかを判別するために、その自己記述レコードを検査する。

【００４２】図５および図６は、接頭部ヘッダ５００（図５）と、一次記憶制御装置４０５によって生成されたレコード・セット情報６００（図６）とを含む、各自己記述レコードごとにＰＤＭ４０４が作成するジャーナル・レコード形式を示している。各自己記述レコードは、それぞれの時間間隔の時間順に二次ＤＡＳＤ４１６に適用できるように、それぞれの時間間隔ごとにさらにＳＤＭ４１４によってジャーナル処理される。

【００４３】ここで図５を参照して説明すると、各レコード・セットの先頭に挿入される接頭部ヘッダ５００は、接頭部ヘッダ５００と、各レコード・セットごとにＳＤＭ４１４に送信される実際の一次レコード・セット情報６００との長さの合計を記述するための総データ長５０１を含む。操作タイム・スタンプ５０２は、ＰＤＭ４０４が現在処理している操作セットの開始時間を示すタイム・スタンプである。この操作タイム・スタンプ５０２は、１組の一次記憶制御装置４０５に対してＲＥＡＤ　ＲＥＣＯＲＤ　ＳＥＴ機能を実行する際に（シスプレックス・クロック４０７に応じて）ＰＤＭ４０４によって生成される。一次ＤＡＳＤ４０６の書込みの入出力時間６１０（図６）は、各一次記憶制御装置４０５のＲＥＡＤ　ＲＥＣＯＲＤ　ＳＥＴごとに固有のものである。操作タイム・スタンプ５０２は、すべての記憶制御装置で共通のものである。

【００４４】ＲＥＡＤ　ＲＥＣＯＲＤ　ＳＥＴコマンドは、ＰＤＭ４０４によって出されるが、以下の条件のいずれかの場合に予測できる。

（１）一次記憶制御装置４０５の所定のしきい値に基づく、その一次記憶制御装置のアテンション割込み

（２）所定の時間間隔に基づく、一次プロセッサ４０１のタイマ割込み

（３）レコード・セット情報が、使用可能であるがまだ読み取られていない未解決のレコード・セットに関する追加情報を示す場合

条件（２）では、タイマ間隔を使用して、低レベル活動の期間中に二次システム４３１がどの程度遅れて実行するかを制御する。条件（３）は、ＰＤＭ４０４が一次記憶制御装置４０５の活動に遅れないようにするためにさらに活動を駆動する処理間隔中に、ＰＤＭ４０４がすべてのレコード・セットを待ち行列処理しなかった場合に発生する。

【００４５】時間間隔グループ番号５０３は、現行レコード・セット（整合性グループのうちの所与の時間間隔グループについてすべての一次記憶制御装置４０５にわたるレコードのセット）が属す時間間隔（操作タイム・スタンプ５０２とレコード読取り時間５０７によって境界が示される）を識別するためにＰＤＭ４０４が出力する。グループ内順序番号５０４は、所与の時間間隔グループ５０３内の各レコード・セットごとに一次記憶制御装置４０５用のアプリケーションＷＲＩＴＥ入出力の順序を（ＰＤＭ４０４に対して）識別するためにハードウェアが提供するＩＤに基づいて導出される。一次ＳＳＩＤ（補助記憶域ＩＤ）５０５は、各レコード・セットごとに一次記憶制御装置４０５の特定の一次記憶制御装置を明確に識別するものである。二次ターゲット・ボリューム５０６は、パフォーマンス上の考慮事項に応じて、ＰＤＭ４０４またはＳＤＭ４１４のいずれかによって割り当てられる。レコード読取り時間５０７は、すべての一次記憶制御装置４０５に共通の操作タイム・スタンプを提供し、現行間隔のレコード・セットの終了時間を示す。

【００４６】操作タイム・スタンプ５０２およびレコード読取り時間５０７は、各一次記憶制御装置４０５から得た複数組の読取りレコード・セットをグループ分けす

るためにＰＤＭ４０４が使用する。複数組の読取りレコード・セットをグループ分けするための同期はＰＤＭ４０４にだけ重要であるため、ＰＤＭ４０４は、シスプレックス・クロック４０７に接続されていないＰＤＭ４０４だけを動作させる中央処理装置（ＣＰＵ）クロックに同期してもよい。ＰＤＭ４０４はレコード更新を書き込まないが、前述の通り、レコード更新は共通の時間源に同期していなければならない。

【００４７】次に図６を参照して説明すると、レコード・セット情報６００は、一次記憶制御装置４０５によって生成され、ＰＤＭ４０４によって収集される。更新固有情報６０１～６１０は、レコード更新が行われた実際の一次ＤＡＳＤ４０６を含む各レコードの一次装置ユニット・アドレス６０１を含む。シリンダ番号／ヘッド番号（ＣＣＨＨ）６０２は、各レコード更新ごとの一次ＤＡＳＤ４０６上の位置を示す。一次記憶制御装置のセッションＩＤである一次ＳＳＩＤ６０３は、一次ＳＳＩＤ５０５と同じものである。状況フラグ６０４は、特定のデータ・レコード６２０が後に続くかどうかに関する状況情報を提供する。順序番号６０５および６３０は、レコード・セット全体（ＰＤＭ４０４に転送されたすべてのデータ）が読み取られたかどうかを示すために各レコードに番号を１つずつ割り当てる。一次ＤＡＳＤ書込み入出力タイプ６０６は、各レコードについて行われた書込み操作のタイプを識別する操作標識である。この操作標識は、更新書込み、フォーマット書込み、部分トラック・レコード・フォロー、完全トラック・データ・フォロー、消去コマンド実行、または全書込み実行を含む。検索引数６０７は、最初に読み取られたレコード・セット・データ・レコード６２０に関する初期位置決め情報を示す。セクタ番号６０８は、レコードが更新されたセクタを識別する。カウント・フィールド６０９は、後続の特定のレコード・データ・フィールド６２０の数を記述する。一次ＤＡＳＤ４０６の書込み更新が行われたホスト・アプリケーション時間は、更新時間６１０に記録される。特定のレコード・データ６２０は、各レコード更新ごとのカウント／キー／データ（ＣＫＤ）フィールドを提供する。最後に、順序番号６３０は、読み取られたレコード・セット全体がＰＤＭ４０４に転送されたかどうかを示すために順序番号６０５と比較される。

【００４８】一次ＤＡＳＤ４０６でレコード更新が書き込まれたのと同じ順序でＳＤＭ４１４がそのレコード更新をコピーできるように、ソフトウェア・グループが呼び出した整合性グループで更新レコードが処理される。整合性グループを作成するのに使用する情報（すべての記憶制御装置４０５から収集したすべてのレコード・セットにわたる）は、操作タイム・スタンプ５０２、時間間隔グループ番号５０３、グループ内順序番号５０４、一次制御装置ＳＳＩＤ５０５、レコード読取り時間５０７、一次装置アドレス６０１、一次ＳＳＩＤ６０３、および状況フラグ６０４を含む。１つの時間間隔グループ用のすべてのレコードがＳＤＭ４１４側で各記憶制御装置４０５ごとに受け取られたかどうかを判別するのに使用する情報は、時間間隔グループ番号５０３、グループ内順序番号５０４、物理制御装置ＩＤ５０５、および一次ＳＳＩＤ６０３および各操作時間間隔ごとに各一次記憶制御装置４０５から返される読取りレコード・セットの総数を含む。完全復旧可能な一次ＤＡＳＤ４０６のレコード更新と同等に二次ＤＡＳＤ４１６上にレコード更新を配置するのに必要な情報は、二次ターゲット・ボリューム５０６、ＣＣＨＨ６０２、一次ＤＡＳＤ書込み入出力タイプ６０６、検索引数６０７、セクタ番号６０８、カウント６０９、更新時間６１０、および特定のレコード・データ６２０を含む。

【００４９】図７および図８は、復旧時間とジャーナル転送時間を単純化した現行ジャーナル内容を記述するための状態テーブル７００とマスタ・ジャーナル８００をそれぞれ示す。状態テーブル７００は、ＰＤＭ４０４とＳＤＭ４１４が収集し、両者に共通の構成情報を提供し、一次記憶制御装置のセッションＩＤ（ＳＳＩＤ番号）およびその制御装置でのボリュームと、対応する二次記憶制御装置のセッションＩＤおよび対応するボリュームとを含む。このため、構成情報は、どの一次ボリューム７１０または一次ＤＡＳＤエクステントが二次ボリューム７１１または二次ＤＡＳＤエクステントにマッピングされるかを追跡する。状態テーブル７００まで単純に拡張して部分ボリューム・エクステント７１２（ＣＣＨＨからＣＣＨＨまで）を示す場合、部分ボリューム遠隔コピーは、ここに記載するのと同じ非同期遠隔コピー方法を使用して達成できるが、完全ボリュームの場合より細分性（トラックまたはエクステント）はより細かくなる。

【００５０】マスタ・ジャーナル８００は、整合性グループ番号、ジャーナル・ボリューム上の位置、および操作タイム・スタンプを含む。また、マスタ・ジャーナル８００は、整合性グループにグループ化した特定のレコード更新を維持する。状態テーブル７００とマスタ・ジャーナル８００は、災害復旧をサポートするため、一次システム４１０がもはや存在しないスタンドアロン環境で動作できなければならない。

【００５１】制御項目全体が正しく書き込まれるようにするため、タイム・スタンプ制御は各マスタ・ジャーナル８００の前後に置かれる。このタイム・スタンプ制御は、さらに二次ＤＡＳＤ４１７に書き込まれる。制御要素は二重項目（１）および（２）を含み、次に示す例のように一方の項目が必ず現行項目になる。

（１）タイム・スタンプ制御｜制御情報｜タイム・スタンプ制御

（２）タイム・スタンプ制御｜制御情報｜タイム・スタンプ制御

いかなる時点でも、（１）または（２）のいずれかの項目が現行または有効項目になるが、有効項目は前後に等しいタイム・スタンプ制御を持つ項目である。災害復旧では、制御情報を得るために、最新のタイム・スタンプを持つ有効項目を使用する。この制御情報は、状態情報（記憶制御装置、装置、および適用される整合性グループに関する環境情報）とともに、二次記憶制御装置４１５にどのレコード更新が適用されたかを判別するのに使用する。

【００５２】整合性グループ

所定の時間間隔の間にすべての一次記憶制御装置４０５にわたるすべての読取りレコード・セットが二次側４３１で受け取られると、ＳＤＭ４１４は、受け取った制御情報を解釈し、レコード更新が最初に一次ＤＡＳＤ４０６上で書き込まれたのと同じ順序でそのレコード更新が適用されるように、受け取った読取りレコード・セットをレコード更新・グループとして二次ＤＡＳＤ４１６に適用する。このため、一次側アプリケーションの順序（データ保全性）整合性はすべて二次側４３１で維持される。このプロセスは、以下、整合性グループの形成と呼ぶ。整合性グループの形成は、次のような仮定に基づいて行われる。（Ａ）独立しているアプリケーション書込みが制御装置の順序命令に違反しない場合は、どのような順序でもアプリケーション書込みを実行できる。（Ｂ）従属しているアプリケーション書込みは、タイム・スタンプの順に実行しなければならないため、アプリケーションは、書込み番号１から制御装置終了、装置終了を受け取る前に従属書込み番号２を実行することができない。（Ｃ）第二の書込みは必ず（１）遅いタイム・スタンプを持つ第一の書込みと同じレコード・セット整合性グループに入るか、（２）後続のレコード・セット整合性グループに入る。

【００５３】図９を参照して説明すると、同図には、記憶制御装置ＳＳＩＤ１、ＳＳＩＤ２、およびＳＳＩＤ３など（記憶制御装置はいくつでも含めることができるが、この例では明解にするため３つ使用する）に関する整合性グループの形成例（整合性グループは一次側４２１または二次側４３１のいずれにも形成できるはずである）が示されている。時間間隔Ｔ１、Ｔ２、およびＴ３は昇順に発生するものと想定する。時間間隔Ｔ１の操作タイム・スタンプ５０２は、記憶制御装置ＳＳＩＤ１、ＳＳＩＤ２、およびＳＳＩＤ３について設定されている。ＰＤＭ４０４は、時間間隔Ｔ１～Ｔ３の間に記憶制御装置ＳＳＩＤ１、２、および３からレコード・セット・データを入手する。時間間隔Ｔ１のＳＳＩＤ１、２、および３に関するレコード・セットは、時間間隔グループ１であるＧ１（時間間隔グループ番号５０３）に割り当てられる。グループ内順序番号５０４は、ＳＳＩＤ１、２、および３のそれぞれについて示され、この場合、ＳＳＩＤ１は１１：５９、１２：００、および１２：０１に３つの更新を持ち、ＳＳＩＤ２は１２：００および１２：０２に２つの更新を持ち、ＳＳＩＤ３は１１：５８、１１：５９、および１２：０２に３つの更新を持つ。時間間隔Ｔ２およびＴ３のレコード・セットは列挙されているが、簡略化のため、更新時間の例は示されていない。

【００５４】ここで、二次側４３１で受け取った制御情報およびレコード更新に基づいて、整合性グループＮを生成することができる。時間間隔グループ番号１のレコード更新が時間間隔グループ番号２のレコード更新より遅くならないようにするため、記憶制御装置ＳＳＩＤ１、２、および３のそれぞれの最後のレコード更新の最も早い読取りレコード・セット時間と等しい、最小時間が設定される。この例では、最小時間は１２：０１になる。最小時間と等しいかそれ以上の読取りレコード・セット時間を有するレコード更新はすべて整合性グループＮ＋１に含まれる。１つのボリュームに対する２つのレコード更新時間が等しい場合、シスプレックス・クロック４０７の十分な解像度が与えられる可能性はほとんどないが、時間間隔グループＮ内の早い順序番号を持つレコード更新は、整合性グループＮ用のそのグループとともに保管される。ここで、レコード更新は、読取りレコード・セット時間に基づいて順序づけされる。複数のレコード更新の時間が等しい場合、小さい順序番号を持つレコード更新は、大きい順序番号を持つレコード更新の前に置かれる。これに対して、複数のレコード更新のタイム・スタンプが等しいが、ボリュームが異なる場合は、そのレコード更新が同じ整合性グループに保管されている限り、任意の順序にすることができる。

【００５５】一次記憶制御装置４０５が、指定の時間間隔の間に読取りレコード・セットへの応答を完了しなかった場合、その一次記憶制御装置４０５が完了するまで、整合性グループを形成することはできない。一次記憶制御装置４０５がその操作を完了しなかった場合は、未着割込みのために、システムの未着割込みハンドラが制御権を受け取り、操作が終了する。これに対して、一次記憶制御装置４０５が適切な時間に操作を完了した場合は、入出力が完了に至り、通常操作が続行される。整合性グループの形成では、一次記憶制御装置４０５に対する書込み操作にタイム・スタンプが付けられると予想される。しかし、プログラムによっては、タイム・スタンプが付けられずに書込みが生成されるものもある。この場合、一次記憶制御装置４０５は、タイム・スタンプとしてゼロを返す。整合性グループの形成は、データが読み取られたタイム・スタンプに基づいて、タイム・スタンプを持たないこれらのレコードの境界を示すことができる。整合性グループの時間別にレコード更新の境界を容易に示せないほど、タイム・スタンプを持たないレコード更新が一定の時間間隔の間に多数発生した場合、二重ボリュームが同期していないというエラーが発生す

る可能性がある。

【００５６】図１０および図１１は、整合性グループを形成する方法を示す流れ図である。図１０を参照して説明すると、このプロセスは、ステップ１０００から始まり、一次側４２１が、行うべき遠隔データ・シャドーイングを確立する。ステップ１０１０では、シスプレックス・クロック４０７を同期クロック（図４）として使用して、すべてのアプリケーション入出力操作にタイム・スタンプが付けられる。ＰＤＭ４０４は、ステップ１０２０で各一次記憶制御装置４０５との遠隔データ・シャドーイング・セッションを開始するが、このステップは、データまたはレコードがシャドーイングされる一次ボリュームの識別を含む。ステップ１０３０では、各アプリケーションＷＲＩＴＥ入出力操作（図６を参照）ごとに一次記憶制御装置４０５によってレコード・セット情報６００がトラッピングされる。

【００５７】ステップ１０４０は、前述の通り、アテンション・メッセージを含むプロンプト、所定のタイミング間隔、または読取りレコード数増加の通知に応じて、ＰＤＭ４０４が、捕捉したレコード・セット情報６００を各一次記憶制御装置４０５から読み取ることを含む。ステップ１０５０でＰＤＭ４０４がレコード・セットの読取りを開始すると、ＰＤＭ４０４は、各レコード・セットの前に特定のジャーナル・レコード（ジャーナル・レコードは、接頭部ヘッダ５００と、レコード・セット情報６００を含む）を作成するための接頭部ヘッダ５００（図５を参照）を付ける。このジャーナル・レコードには、二次側４３１（または一次側４２１）で整合性グループを形成するのに必要な制御情報（およびレコード）が含まれる。

【００５８】ステップ１０６０では、ＰＤＭ４０４が通信リンク４０８を介して（整合性グループがそこで形成される場合は、同じデータ・ムーバ・システム内で）ＳＤＭ４１４に生成したジャーナル・レコードを送信する。ＳＤＭ４１４は、ステップ１０７０で状態テーブル７００を使用し、データ・シャドーイング・セッション用に確立した各時間間隔グループおよび一次記憶制御装置４０５ごとに、受け取ったレコード更新をグループ番号別および順序番号別に収集する。ステップ１０８０でＳＤＭ４１４は、ジャーナル・レコードを検査し、各時間間隔グループごとにすべてのレコード情報を受け取ったかどうかを判別する。ジャーナル・レコードが不完全な場合は、ステップ１０８５によって、ＳＤＭ４１４はＰＤＭ４０４に必要なレコード・セットを再送信するよう通知する。ＰＤＭ４０４が正しく再送信できない場合は、二重ボリューム対に障害が発生している。ジャーナル・レコードが完全な場合は、ＳＤＭ４１４による整合性グループの形成を含むステップ１０９０が実行される。

【００５９】図１１を参照すると、同図には、整合性グループを形成するためのステップ１０９０（図１０）を表すステップ１１００～１１６０が示されている。整合性グループの形成は、ステップ１１００から始まるが、このステップでは、各ソフトウェア整合性グループが二次ＤＡＳＤ４１７（図４）上のＳＤＭ４１４ジャーナル・ログ（"hardened"）に書き込まれる。ステップ１１１０は、時間間隔グループが完全かどうかを判別するテストを実行する。すなわち、各一時記憶制御装置４０５は、少なくとも１つの読取りレコード・セット・バッファを提示したか、レコード・セット・バッファ内にこのようなレコード更新が置かれていないという確認をＰＤＭ４０４から受け取らなければならず、しかもデータ（またはヌル）を持つすべての読取りレコード・セット・バッファがＳＤＭ４１４によって受け取られていなければならない。時間間隔グループが不完全な場合は、ステップ１１１０は、必要なデータが受け取られるまで、一次記憶制御装置４０５からのレコード・セットの読取りを再試行する。エラーが発生した場合は、特定の１つまたは複数の二重ボリューム対に障害が発生している可能性がある。完全な時間間隔グループを受け取ると、ステップ１１２０は、第一の整合性グループ・ジャーナル・レコードを判別する。この第一（または現行）の整合性グループ・ジャーナル・レコードとは、最も早い操作タイム・スタンプ５０２と、同じ操作タイム・スタンプ５０２を持つすべてのレコードの最も早い更新時間６１０とを含むレコードである。

【００６０】ステップ１１３０は、現行整合性グループ・ジャーナル・レコードに含まれるレコードを検査して、どのレコードがそこに最後に含めるレコードかを判別する（一部のレコードは除去され、次の整合性グループ・ジャーナル・レコードに含まれる）。現行整合性グループ・ジャーナル・レコードの最後のレコードは、各一次記憶制御装置４０５ごとに最大更新時間のうちの最小更新時間（最小時間）として判別される（つまり、各一次記憶制御装置４０５の最後の更新が比較され、これらのうちの最も早いものだけが現行整合性グループ・ジャーナル・レコードに残る）。

【００６１】現行整合性グループ・ジャーナル・レコードに残っているこれらのレコード更新は、ステップ１１４０で、更新時間６１０とグループ内順序番号５０４に応じて順序付けされる。レコード更新を持たない一次記憶制御装置４０５は、整合性グループに関与しない。ステップ１１５０では、現行整合性グループの残っているレコード更新（最小時間より遅い更新時間を持つもの）が、次の整合性グループに渡される。それぞれのグループ内順序番号５０４は、空バッファで終わり、その操作時間間隔の間にすべての読取りレコード・セットが読み取られたことを示すはずである。空バッファがない場合は、現行ソフトウェア整合性グループ内の最後のレコードを定義するステップ１１２０を、レコード読取り時間

５０７および更新時間６１０と併せ使用して、一次記憶制御装置４０５におけるアプリケーションＷＲＩＴＥ入出力操作の正しい順序を決定することができる。

【００６２】ステップ１１６０は、完全災害復旧の制約の下で特定の書込み更新が二次ＤＡＳＤ４１６に適用される、遠隔データ・シャドーイング・プロセスのバックエンドを表している。二次ＤＡＳＤ４１６に更新内容を書き込む際に入出力エラーが発生するか、または二次側４３１全体が停止し、最初期設定された場合は、書込みプロセスに入っていた整合性グループ全体を最初から再適用することができる。このため、どの二次ＤＡＳＤ４１６の入出力が行われたか、どの入出力が行われなかったか、どの入出力が処理中かなどを追跡せずに、遠隔シャドーイングを行うことができる。

【００６３】二次入出力書込み

ステップ１１６０の重要な構成要素は、二次側４３１が一次側４２１から後れをとらないように、ＰＤＭ４１４によってレコードが二次ＤＡＳＤ４１６に効率よく書き込まれることである。必要な効率は、主に、様々な二次ＤＡＳＤ４１６への複数の入出力操作を同時に実行することによって達成される。二次ＤＡＳＤ４１６を一度に１つずつ連続して書き込むと、二次側４３１は一次側４２１からかなり遅れてしまう恐れがある。単一チャネル・コマンド・ワード（ＣＣＷ）連鎖を介して単一の二次装置宛ての整合性グループごとにレコードを書き込めば、二次側４３１ではさらに高い効率が得られる。それぞれの単一ＣＣＷ連鎖内では、そこで行われる各二次ＤＡＳＤ４１６のデータ・トラックへの入出力操作が一次ボリュームでの発生順に維持されている限り、その入出力操作をさらに最適化することができる。

【００６４】特定の整合性グループ用の二次入出力操作を単一ＣＣＷ連鎖内で最適化する場合、一部は一次書込み入出力操作のパターンに基づいて行われ、一部は二次ＤＡＳＤ４１６の物理特性に基づいて行われる。最適化は、二次ＤＡＳＤ４１６がカウント／キー／データ（ＣＫＤ）か、拡張カウント／キー／データ（ＥＣＫＤ）か、固定ブロック方式（ＦＢＡ）かなどに応じて、多少変化する可能性がある。その結果、所与の時間間隔の間に１つの一次ＤＡＳＤ４０６に対して行われる複数のＷＲＩＴＥ入出力（ｍ）は、１つの二次ＤＡＳＤ４１６のボリュームに対する単一のＳＴＡＲＴ入出力操作に削減することができる。このように二次記憶制御装置４１５に対するＳＴＡＲＴ入出力の回数をｍ：１に最適化すると、二次ＤＡＳＤ４１６は後れをとらずに済み、それにより、一次側４２１のレコード更新をもっと精密にシャドーイングすることができる。

【００６５】正常な遠隔データ・シャドーイングとそれによる二次入出力の最適化の重要点は、一貫性のあるコピーを復旧に使用できるように、二次ＤＡＳＤ４１６に対して同時に行う複数の入出力操作のいずれかで発生する回復不能エラーを最小限にすることである。所与の二次書込みで失敗すると、その後の従属書込みが条件付け書込みを伴わずに記録される恐れがある（たとえば、実際にはデータベース用の実際の更新書込みが失敗に終わっているのにデータベース・レコードが更新されたことを示すログ項目は、二次ＤＡＳＤ４１６のコピーの順序整合性に違反する）。

【００６６】その更新失敗が復旧されるまで、失敗した二次ＤＡＳＤ４１６のコピーはアプリケーションの復旧に使用できなくなる。失敗した更新は、ＳＤＭ４１４によってＰＤＭ４０４から現行コピーを要求することで修正できるはずである。その間に二次データ・コピーは不整合になり、そのため、ＰＤＭ４０４が現行更新で応答し、それ以前の他のすべての更新がＰＤＭ４０４によって処理されるまで使用できなくなる。通常、失敗した更新の復旧に要する時間は、十分な災害復旧保護のために受け入れられないほど長い非復旧ウィンドウを示す。

【００６７】効果的な二次側４３１の入出力最適化は、所与の整合性グループについて書き込まれるデータ・レコード・セットを検査し、ＥＣＫＤ方式などの二次ＤＡＳＤ４１６の特定の方式の規則に基づいて連鎖を構築することで実現される。ここに開示する最適化技法は、入出力エラーが発生した場合に整合性グループを適用する際に、ＣＣＷ連鎖を再実行できるように、または、二次初期プログラム・ロード（ＩＰＬ）復旧の場合に、データを紛失せずに整合性グループ全体を再適用できるように、入出力エラーからの復旧を単純化するものである。

【００６８】図１２は、ＥＣＫＤ方式用のすべてのＷＲＩＴＥ入出力の組合せに対応するＣＣＷ連鎖を構築するための完全整合性グループ復旧（ＦＣＧＲ）の規則を要約して示すもので、ここではＣＣＨＨＲレコード形式が使用される（シリンダ番号、ヘッド番号、レコード番号）。図１２は、１つの整合性グループの範囲内でＤＡＳＤトラックに対して行われるＷＲＩＴＥ入出力操作の可能な組合せをそれぞれ検査することによって作成される。図１２のＦＣＧＲ規則（図１４および図１５に記載する）は、整合性グループを適用する際のエラーについて完全復旧を行うためにデータ配置（二次ＤＡＳＤ４１６の入出力書込みＣＣＷ連鎖）を管理する場合に従うものである。図１２に示すＦＣＧＲ規則は、新しいＷＲＩＴＥ入出力操作が追加されるたびに適切に拡張されることになる。これらの規則は、二次側４３１のハードウェアまたはソフトウェアで実現することができる。ＦＣＧＲ規則は、同一ＤＡＳＤトラック分析に対するＲＥＡＤレコード・セットを、一次ＤＡＳＤ４０６のＷＲＩＴＥ入出力タイプ、検索引数、およびカウント・フィールドとキー・フィールドの検査に還元するので好都合である。

【００６９】図１２に示すように整合性グループの書込み操作を検査せずにＤＡＳＤトラックが書き込まれる

と、前に書き込まれたデータ・レコードの再書込みができなくなる可能性がある。たとえば、以下の内容の連鎖があると想定する。

レコード５へのＷＲＩＴＥ　ＵＰＤＡＴＥ

レコード１へのＦＯＲＭＡＴ　ＷＲＩＴＥ

この場合、レコード１とレコード５は同じＤＡＳＤトラックに存在し、レコード１がレコード５の前に置かれる。レコード５はＵＰＤＡＴＥ　ＷＲＩＴＥ　ＣＣＷによって更新されるが、ＦＯＲＭＡＴ　ＷＲＩＴＥ入出力ＣＣＷは、トラックの残りを消去してレコード１を更新するため、レコード５は削除される。この連鎖を再実行しなければならない場合、レコード５の先頭に配置するＬＯＣＡＴＥＲＥＣＯＲＤ　ＣＣＷが位置決めポイントを持たなくなる（レコード５が存在しなくなる）ため、この連鎖は先頭から完全に回復することができない。一次側４２１ですでに書込み操作が正常に行われているので、データの整合性と保全性を維持するには、二次ＤＡＳＤ４１６上の整合性グループ全体をいつでも適用できることが必要である。

【００７０】図１６のステップ１４１０～１４７０は、図１１のステップ１１６０によって表され、図１２に定義されるＦＣＧＲ規則を使用するプロセスの詳細を示すものである。ステップ１４１０でＳＤＭ４１４は、現行整合性グループの各種レコードを２つのカテゴリに分割する。第一のカテゴリは、同じ二次ＤＡＳＤボリューム向けの入出力命令を含み、第二のカテゴリは、第一のカテゴリに含まれるレコードのうち、同じＣＣＨＨ向けのレコードの入出力命令を含む（すなわち、同じＤＡＳＤトラックに更新されるレコード）。

【００７１】現行整合性グループのレコードのカテゴリ分類後、ステップ１４２０では、トラック上のデータ配置を識別し、トラック／レコードのアドレス指定を行うために、アプリケーションのＷＲＩＴＥ入出力およびＳＤＭ４１４のＷＲＩＴＥ入出力を二次ＤＡＳＤ４１６の方式、たとえば、ＥＣＫＤ方式のＦＣＧＲ規則（図１２を参照）に適合させる。ＳＤＭ４１４は、ステップ１４３０で、同じボリュームに対する二次ＤＡＳＤのＷＲＩＴＥ入出力操作を単一入出力ＣＣＷ連鎖にグループ分けする。ステップ１４４０は、実際の二次ＤＡＳＤ４１６の書込み用の検索引数と特定のレコード・データ（ＣＫＤフィールド）に応じて、それぞれの二次ＤＡＳＤ４１６のヘッド・ディスク・アセンブリ（ＨＤＡ）を移動させることを含む。

【００７２】ステップ１４５０では、後続の書込み操作によって、前の書込み操作またはＤＡＳＤ検索引数（ここで消去されるレコードでの分割など）を無効にするかどうかを判別するために図１２のＦＣＧＲ規則を使用して、第二のカテゴリ（通常、レコードを受け取るトラックごとに１つずつ、複数の第二のカテゴリが存在する）を構成するこれらのレコードについてＲＥＡＤ　ＳＥＴ　ＢＵＦＦＥＲＳ１と２を比較する。ＲＥＡＤ　ＳＥＴ　ＢＵＦＦＥＲＳ１と２は隣接する読取りレコード・セットが含まれている。ＦＣＧＲ規則に従うと、エラーが発生した場合に、一次側４２１からレコード更新を再度受け取らなくても、ＳＤＭ４１４は整合性グループ全体を再書込みできるようになる。ＳＤＭ４１４が現行整合性グループを二次ＤＡＳＤ４１６に適用した後、ステップ１４６０では、状態テーブル（図７）とマスタ・ジャーナル（図８）を更新する。

【００７３】ステップ１４７０は、次の整合性グループ（現行整合性グループになるもの）を獲得して、処理をステップ１４１０に戻すので、遠隔コピー・プロセスはリアルタイムで続行される。一次側４２１から二次側４３１への通信が終了した場合、遠隔コピー・プロセスは停止する。この通信は、ボリューム対がＰＤＭ４０４によってプロセスから削除された場合、一次側が破壊された場合（災害が発生した場合）、規則的な運転停止が行われた場合、または二次側４３１で特定の引継ぎ処置が行われた場合に終了することがある。二次側４３１でジャーナル処理された整合性グループは、引継ぎ操作中に二次ＤＡＳＤ４１６に適用することができる。一次側４２１が捕捉したデータのうち、ＳＤＭ４１４によって完全に受け取られていないデータだけが紛失する。

【００７４】要約すると、これまで同期および非同期遠隔データ二重化システムについて説明してきた。非同期遠隔データ二重化システムは、記憶域ベースのリアルタイム・データ・シャドーイングを提供する。一次側は、レコード更新を生成するアプリケーションを実行し、一次側から離れた位置にある二次側は、レコード更新をシャドーイングして、一次側の災害復旧を行う。この非同期遠隔データ二重化システムは、一次側の時間依存プロセスを同期させるためのシスプレックス・クロックと、アプリケーションを実行するための一次側の一次プロセッサとを含み、一次プロセッサはそこに一次データ・ムーバを有している。一次プロセッサには、各レコード更新ごとに書込み入出力操作を出すために複数の一次記憶制御装置が連結され、それぞれの一次記憶制御装置ＤＡＳＤ書込み入出力操作がシスプレックス・クロックに同期している。複数の一次記憶装置はこの書込み入出力操作を受け取り、それに応じてレコード更新をそこに格納する。一次データ・ムーバは、各レコード更新ごとに複数の一次記憶制御装置からレコード・セット情報を収集し、所定のグループのレコード・セット情報に接頭部ヘッダを付加する。この接頭部ヘッダと所定のグループのレコード・セット情報が自己記述レコード・セットを形成する。それぞれのレコード・セット情報は、一次装置アドレス、シリンダ番号とヘッド番号（ＣＣＨＨ）、レコード更新順序番号、書込み入出力タイプ、検索引数、セクタ番号、およびレコード更新時間を含む。接頭部ヘッダは、総データ長、操作タイム・スタンプ、時間間隔

グループ番号、およびレコード読取り時間を含む。二次側の二次プロセッサは二次データ・ムーバを有し、この二次データ・ムーバが一次側から自己記述レコード・セットを受け取る。二次プロセッサには複数の二次記憶制御装置が連結され、二次記憶制御装置にはレコード・更新のコピーを格納するために複数の二次記憶装置が連結されている。二次データ・ムーバは、送られてきた自己記述レコード・セットが完全なものであるかどうかを判定し、その自己記述レコード・セットから整合性グループを形成し、さらに、レコード更新が複数の一次記憶装置に書き込まれた順序と整合する順序で複数の二次記憶装置に書き込むために、各整合性グループから得たレコード更新を複数の二次記憶制御装置に出力する。

【００７５】特に本発明の実施例に関連して本発明を図示し説明してきたが、当業者は、本発明の精神および範囲を逸脱せずに形態および詳細の様々な変更が可能であることに留意されたい。たとえば、整合性グループは、受け取った自己記述レコード・セットに基づいて二次データ・ムーバによって形成されたものとして説明してきたが、書込みレコード・セットに基づいて一次側で整合性グループを形成するか、二次側の他の部分で形成することもできる。一次側と二次側の記憶装置の形式は、同じである必要はない。たとえば、ＣＫＤレコードを固定ブロック方式（ＦＢＡ）タイプのレコードなどに変換することも可能である。また、記憶装置は、ＤＡＳＤ装置に限定されているわけではない。

【００７６】

【発明の効果】本発明の実施により、災害復旧のためにＤＡＳＤデータを二次側にシャドーイングするための改良された設計および方法を提供することができる。

【００７７】まとめとして、本発明の構成に関して以下の事項を開示する。

【００７８】（１）一次プロセッサで実行される１つまたは複数のアプリケーションによって生成されたデータ更新が一次記憶サブシステムによって受け取られ、一次記憶サブシステムは入出力書込み操作によって各データ更新を書き込まれ、各書込み入出力操作にタイム・スタンプが付けられ、共通タイマによってタイム・スタンプの同期が取られ、一次プロセッサと同じ地域にあるか一次プロセッサから離れた位置にあるかにかかわらず、二次システムは、災害復旧のために二次側が使用できるように順序整合性のある順序でデータ更新をシャドーイングする、災害復旧機能を提供するために整合性グループを形成する方法において、前記方法が、（ａ）一次記憶サブシステムで発生する各書込み入出力操作にタイム・スタンプを付けるステップと、（ｂ）各データ更新ごとに一次記憶サブシステムから書込み入出力操作レコード・セット情報を収集するステップと、（ｃ）データ更新とそれぞれのレコード・セット情報から自己記述レコード・セットを生成し、この自己記述レコード・セットが、二次システムだけによる書込み入出力操作の順序を再生成するために十分な制御情報を含んでいるステップと、（ｄ）自己記述レコード・セットを間隔グループにグループ分けし、各間隔グループが、操作タイム・スタンプ開始時間から測定され、所定の間隔しきい値の間持続するステップと、（ｅ）最も早い操作タイム・スタンプを有する自己記述レコード・セットの間隔グループとして現行整合性グループを選択し、一次記憶サブシステムでの入出力書込み操作の時間順に基づいて、個々のデータ更新が現行整合性グループ内で順序付けされるステップとを含む方法。

（２）各間隔グループの開始時間を識別するための操作タイム・スタンプに基づいて、一次記憶サブシステムとのセッションを開始し、各間隔グループの境界が連続する操作タイム・スタンプによって示されることが、ステップ（ｂ）にさらに含まれることを特徴とする、上記（１）に記載の方法。

（３）各間隔グループを記述する接頭部ヘッダを追加することがステップ（ｄ）に含まれることを特徴とする、上記（１）に記載の方法。

（４）自己記述レコード・セットからなる間隔グループを二次側に送信するステップ（ｆ）をさらに含むことを特徴とする、上記（３）に記載の方法。

（５）受け取った各自己記述レコード・セットが完全なものであるかどうかを二次側で判定するステップ（ｇ）をさらに含むことを特徴とする、上記（４）に記載の方法。

（６）自己記述レコード・セットが不完全であると二次側が判定した場合に、欠落データ更新を再送信するよう二次側が一次側に要求することが、ステップ（ｇ）にさらに含まれることを特徴とする、上記（５）に記載の方法。

（７）各時間間隔グループが完全なものであるかどうかを二次側で判定するステップ（ｈ）をさらに含むことを特徴とする、上記（６）に記載の方法。

（８）間隔グループが不完全であると二次側が判定した場合に、欠落レコード・セットを再送信するよう二次側が一次側に要求することが、ステップ（ｈ）にさらに含まれることを特徴とする、上記（７）に記載の方法。

（９）整合性グループで順序づけされたように、対応する一次側の書込み入出力操作の順序に応じて、二次側で受け取ったデータ更新を二次記憶サブシステムに書き込むステップ（ｉ）をさらに含むことを特徴とする、上記（８）に記載の方法。

（１０）災害復旧のために遠隔データ・シャドーイングを行うシステムにおいて、このシステムは、一次データ・ムーバと、レコード更新を生成するアプリケーションとを実行する一次プロセッサを有する一次側を含み、一次プロセッサは、一次プロセッサから一次記憶サブシステムに出される書込み入出力操作に応じてレコード更新

を格納するための記憶装置を有する一次記憶サブシステムに連結され、一次側は、一次側の時間依存操作を同期させるために共通のシステム・タイマをさらに含み、システムは、一次プロセッサと通信する二次プロセッサと、順序整合性のある順序でレコード更新のコピーを格納するための二次記憶サブシステムとを有する二次側をさらに含み、（ａ）一次記憶サブシステムの各書込み入出力操作にタイム・スタンプを付けるステップと、（ｂ）一次記憶サブシステム内の各記憶装置とのセッションを確立するステップと、（ｃ）一次記憶サブシステム内の各記憶装置からレコード・セット情報を収集するステップと、（ｄ）レコード・セットとそれぞれのレコード・セット情報を一次データ・ムーバに読み込むステップと、（ｅ）各レコード・セットの前にヘッダを付け、それに基づく自己記述レコード・セットを生成するステップと、（ｆ）所定の時間間隔に応じて、時間間隔グループ単位で自己記述レコード・セットを二次プロセッサに送信するステップと、（ｇ）自己記述レコード・セットから整合性グループを形成するステップと、（ｈ）順序整合性のある順序で各整合性グループのレコード更新を二次記憶サブシステムにシャドーイングするステップとを含む、順序整合性のある順序でレコード更新をシャドーイングするための方法。

（１１）レコード・セットが非同期的に二次プロセッサに送信されることを特徴とする、上記（１０）に記載の方法。

（１２）ステップ（ｇ）が二次側で行われることを特徴とする、上記（１０）に記載の方法。

（１３）受け取った各自己記述レコード・セットが完全であるかどうかを二次側で判定するステップが、ステップ（ｆ）にさらに含まれることを特徴とする、上記（１０）に記載の方法。

（１４）受け取った自己記述レコード・セットが不完全であると二次側が判定した場合に、欠落レコード更新を再送信するよう一次側に要求するステップが、ステップ（ｆ）にさらに含まれることを特徴とする、上記（１３）に記載の方法。

（１５）各時間間隔グループが完全であるかどうかを二次側で判定するステップ（ｉ）をさらに含むことを特徴とする、上記（１０）に記載の方法。

（１６）時間間隔グループが不完全であると二次側が判定した場合に、欠落レコード・セットを再送信するよう一次側に要求することが、ステップ（ｉ）にさらに含まれることを特徴とする、上記（１５）に記載の方法。

（１７）ステップ（ｃ）が、レコード・セット情報において、各レコード更新が格納されている一次記憶装置上の物理的位置を識別することを特徴とする、上記（１０）に記載の方法。

（１８）ステップ（ｃ）が、レコード・セット情報において、セッション内に一次記憶装置に格納された各レコード更新の順序と更新時間を識別することを特徴とする、上記（１７）に記載の方法。

（１９）ステップ（ｅ）が、接頭部ヘッダにおいて、セッション用の間隔グループ番号と、そこで参照される各レコード更新用のグループ内順序を識別することを特徴とする、上記（１０）に記載の方法。

（２０）１つまたは複数のアプリケーションを実行する一次プロセッサを有し、この１つまたは複数のアプリケーションがレコード更新を生成し、一次プロセッサがそれに基づく自己記述レコード・セットを生成し、自己記述レコード・セットが二次システムに送られ、二次システムがリアルタイム災害復旧のために自己記述レコード・セットに基づいて順序整合性のある順序でレコード更新をシャドーイングし、一次プロセッサが一次記憶サブシステムに連結され、一次記憶サブシステムがレコード更新を受け取って、そこに各レコード更新を格納するために書込み入出力操作を実行し、一次プロセッサが、同期を取るためにアプリケーションと一次記憶サブシステムに共通の時間源を提供するためのシスプレックス・クロックと、各レコード更新ごとにレコード・セット情報を提供するよう一次記憶サブシステムに指示し、複数のレコード更新と、それに対応するそれぞれのレコード・セット情報を時間間隔グループにグループ分けし、それに接頭部ヘッダを挿入し、各時間間隔グループがレコード・セットを自己記述する一次データ・ムーバ手段とを含む一次システム。

（２１）一次記憶サブシステムが、書込み入出力操作を出す複数の一次記憶制御装置と、複数の一次記憶サブシステムに連結された複数の一次記憶装置とを含むことを特徴とする、上記（２０）に記載の一次システム。

（２２）複数の一次記憶装置が直接アクセス記憶装置であることを特徴とする、上記（２１）に記載の一次システム。

（２３）一次データ・ムーバ手段が、各時間間隔グループに関与する複数の一次記憶制御装置のうちの各一次記憶制御装置用のそれぞれの書込み入出力操作ごとにレコード・セット情報を収集することを特徴とする、上記（２２）に記載の一次システム。

（２４）一次プロセッサにおいて、各書込み入出力操作にシスプレックス・クロックに関するタイム・スタンプが付けられ、各書込み入出力操作が、複数の一次記憶制御装置のうちの１つの一次記憶制御装置に出され、各一次記憶制御装置が、タイム・スタンプを保持し、対応する読取りレコード・セットに入れてそのタイム・スタンプを一次データ・ムーバ手段に返すことを特徴とする、上記（２３）に記載の一次システム。

（２５）各レコード・セット情報が、複数の一次記憶装置のうちの１つの一次記憶装置上における対応するレコード更新の物理的位置を識別することを特徴とする、上記（２４）に記載の一次システム。

（２６）各レコード・セット情報が、対応するレコード更新の一次サブシステムＩＤ、一次装置アドレス、シリンダ番号、およびヘッド番号を識別することを特徴とする、上記（２４）に記載の一次システム。

（２７）一次データ・ムーバ手段が、１つの時間間隔グループに関与するすべての一次記憶制御装置にわたる各書込み入出力更新の相対順序を識別することを特徴とする、上記（２４）に記載の一次システム。

（２８）レコード更新をジャーナル処理し、一次システムおよび二次システム上の各レコード更新の記憶場所を相互参照するために、一次データ・ムーバ手段が状態テーブルを作成することを特徴とする、上記（２７）に記載の一次システム。

（２９）一次データ・ムーバ手段が、二次システムに状態テーブルを送信することを特徴とする、上記（２７）に記載の一次システム。

（３０）一次側と二次側を含み、二次側が災害復旧のために一次側のレコード更新をリアルタイムでシャドーイングし、レコード更新が一次側で実行されるアプリケーションによって生成され、一次側が、シスプレックス・クロックと、レコード更新を生成するアプリケーションを実行し、各レコード更新ごとに対応する書込み入出力操作を出し、一次データ・ムーバをそこに有する一次プロセッサと、レコード更新を格納するよう指示され、各レコード更新ごとに出された書込み入出力操作を実行する複数の一次記憶制御装置と、対応する書込み入出力操作に応じて、レコード更新を受け取ってそこに格納する複数の一次記憶装置とを含み、書込み入出力操作が互いに正しい順序で並べられるように、シスプレックス・クロックによって同期が取られた通りに、一次プロセッサによって一次プロセッサと各書込み入出力にタイム・スタンプが付けられ、一次データ・ムーバが、複数組のレコード更新を収集し、複数の一次記憶制御装置のうちのそれぞれによって提供された各レコード・セット情報と対応するレコード更新を組み合わせ、各レコード・セット情報が、それぞれの対応する書込み入出力操作の相対順序と時間を含み、一次データ・ムーバが時間間隔グループ別にレコード更新を収集して、各時間間隔グループに接頭部ヘッダを挿入し、接頭部ヘッダが各時間間隔グループに含まれるレコード更新を識別する情報を含み、各レコード・セット情報と接頭部ヘッダが、自己記述レコード・セットを生成するために組み合わされ、自己記述レコード・セットが二次側に送信され、自己記述レコード・セットが、一次側からの追加の通信がなくても順序整合性のある順序でレコード更新をそこにシャドーイングするために十分な情報を二次側に提供する、遠隔データ・シャドーイング・システム。

（３１）遠隔データ・シャドーイングに関与するために識別されたすべての一次記憶制御装置とのセッションを確立することによって、一次データ・ムーバが時間間隔グループを形成することを特徴とする、上記（３０）に記載の遠隔データ・シャドーイング・システム。

（３２）一次データ・ムーバが、識別されたすべての一次記憶制御装置からレコード・セット情報を収集することを特徴とする、上記（３１）に記載の遠隔データ・シャドーイング・システム。

（３３）一次プロセッサが、二次側に自己記述レコードを送信することを特徴とする、上記（３２）に記載の遠隔データ・シャドーイング・システム。

（３４）一次データ・ムーバが、自己記述レコードから整合性グループを形成することを特徴とする、上記（３２）に記載の遠隔データ・シャドーイング・システム。

（３５）二次側が、送信された自己記述レコードから整合性グループを形成することを特徴とする、上記（３３）に記載の遠隔データ・シャドーイング・システム。

（３６）レコード更新をジャーナル処理し、一次システムおよび二次システム上の各レコード更新の記憶場所を相互参照する状態テーブルを、一次データ・ムーバが作成することを特徴とする、上記（３２）に記載の遠隔データ・シャドーイング・システム。

（３７）複数の一次記憶装置が直接アクセス記憶装置（ＤＡＳＤ）であることを特徴とする、上記（３２）に記載の遠隔データ・シャドーイング・システム。

（３８）各レコード・セット情報が、一次装置アドレスと、シリンダ番号およびヘッド番号（ＣＣＨＨ）と、レコード更新順序番号と、書込み入出力タイプと、検索引数と、セクタ番号と、レコード更新時間とを含むことを特徴とする、上記（３７）に記載の遠隔データ・シャドーイング・システム。

（３９）接頭部ヘッダが、総データ長と、操作タイム・スタンプと、時間間隔グループ番号と、レコード読取り時間とを含むことを特徴とする、上記（３７）に記載の遠隔データ・シャドーイング・システム。

（４０）レコード更新を生成するアプリケーションを実行する一次側を含み、一次側から離れた位置に二次側を有し、二次側がレコード更新をシャドーイングして、一次側に災害復旧を提供する、記憶域ベースのリアルタイム・データ・シャドーイングを行う非同期遠隔データ二重化システムにおいて、非同期遠隔データ二重化システムが、一次側の時間依存プロセスを同期させるためのシスプレックス・クロックと、アプリケーションを実行し、対応するレコード更新用の書込み入出力操作を出し、一次データ・ムーバをそこに有する、一次側の一次プロセッサと、各レコード更新ごとに書込み入出力操作を１つずつ受け取り、それぞれの一次記憶制御装置書込み入出力操作が一次プロセッサによってシスプレックス・クロックと同期している、複数の一次記憶制御装置と、対応する書込み入出力操作に応じて、レコード更新をそこに格納するための複数の一次記憶装置とを含み、一次データ・ムーバが、各レコード更新ごとに複数の一

次記憶制御装置からレコード・セット情報を収集して、所定のグループのレコード・セット情報に接頭部ヘッダを付加し、接頭部ヘッダと所定のレコード・セット情報グループが自己記述レコード・セットを形成し、各レコード・セット情報が、一次装置アドレス、シリンダ番号およびヘッド番号（ＣＣＨＨ）、レコード更新順序番号、書込み入出力タイプ、検索引数、セクタ番号、およびレコード更新番号を含み、接頭部ヘッダが、総データ長、操作タイム・スタンプ、時間間隔グループ番号、およびレコード読取り時間を含み、二次データ・ムーバを有し、その二次データ・ムーバが一次側から自己記述レコード・セットを受け取る、二次側の二次プロセッサと、二次プロセッサに連結された複数の二次記憶制御装置と、レコード更新を格納する複数の二次記憶装置とをさらに含み、二次データ・ムーバが、送信された自己記述レコード・セットが完全なものであるかどうかを判定し、自己記述レコード・セットから整合性グループを形成し、複数の一次記憶装置にレコード更新が書き込まれたときの順序に整合する順序で複数の二次記憶装置に書き込むために各整合性グループから得たレコード更新を複数の二次記憶制御装置に出力する、非同期遠隔データ二重化システム。

【図面の簡単な説明】

【図１】同期遠隔データ・シャドーイング機能を有する災害復旧システムのブロック図である。

【図２】図１の災害復旧システムにより同期遠隔コピーを提供する方法を示す流れ図である。

【図３】入出力エラー回復プログラム（入出力ＥＲＰ）操作の方法を示す流れ図である。

【図４】非同期遠隔データ・シャドーイング機能を有する災害復旧システムのブロック図である。

【図５】図４の一次側からの読取りレコード・セットの前に付く接頭部ヘッダを示すデータ形式図である。

【図６】読取りレコード・セットを構成する各種フィールドを示すデータ形式図である。

【図７】ボリューム構成情報を識別する状態テーブルである。

【図８】図４の二次側が使用するマスタ・ジャーナルである。

【図９】整合性グループを形成するためのシーケンス例である。

【図１０】整合性グループを形成するために情報および読取りレコード・セットを収集する方法を示す流れ図である。

【図１１】整合性グループを形成する方法を示す流れ図である。

【図１２】ＤＡＳＤトラックに対する所与の入出力操作シーケンスの場合のＥＣＫＤ方式装置用の完全整合性グループ復旧規則アプリケーションを示すテーブルである。

【図１３】図１２のテーブルで使用する規則の説明の構成を示す図である。

【図１４】図１２のテーブルで使用する規則の説明の一部である。

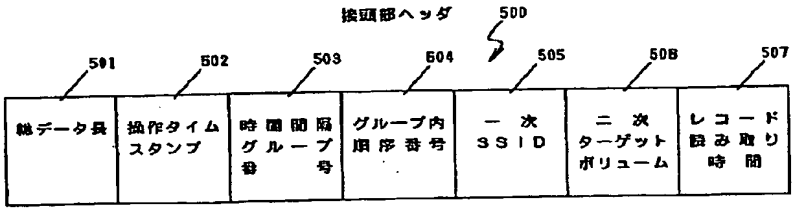
【図１５】図１２のテーブルで使用する規則の説明の一部である。

【図１６】完全整合性グループ復旧機能を持つ二次側に読取りレコード・セット・コピーを書き込む方法の流れ図である。

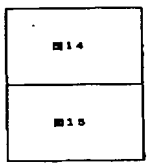
【符号の説明】

１　一次プロセッサ
２　入出力ＥＲＰ
３　一次記憶制御装置
４　一次ＤＡＳＤ
５　二次プロセッサ
６　二次記憶制御装置
７　二次ＤＡＳＤ
８　対等通信リンク
９　エンタープライズ・システム接続（ＥＳＣＯＮ）リンク
１０　災害復旧システム
１１　ホスト間通信リンク
１２　チャネル
１３　チャネル
１４　一次側
１５　二次側

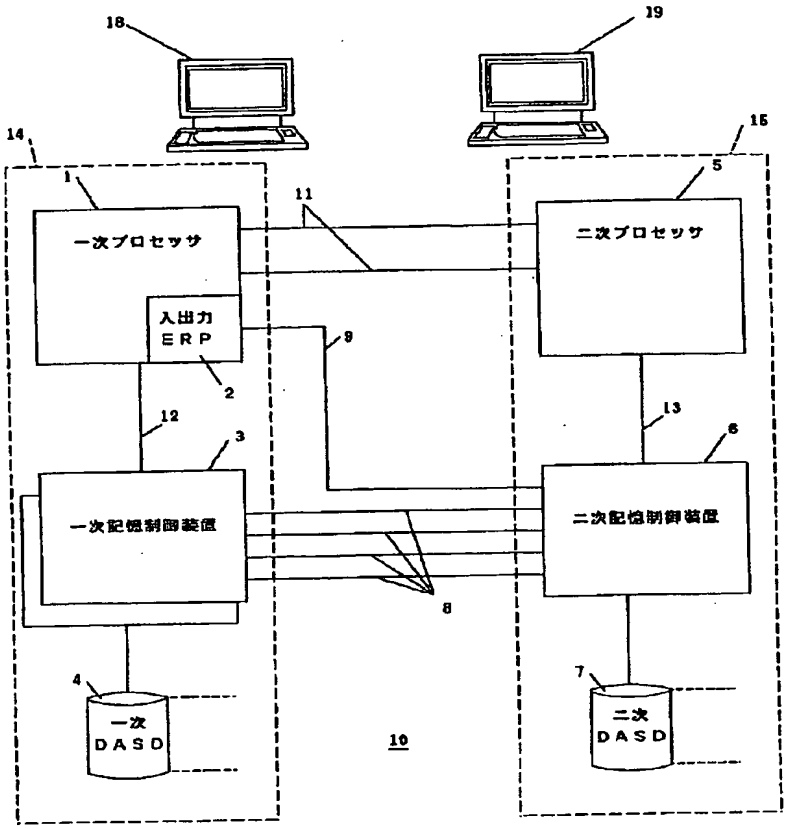
【図５】



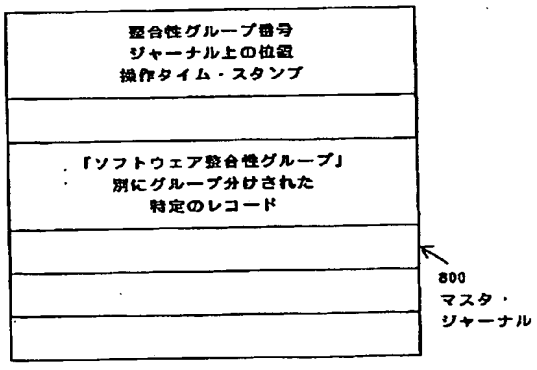
【図１３】



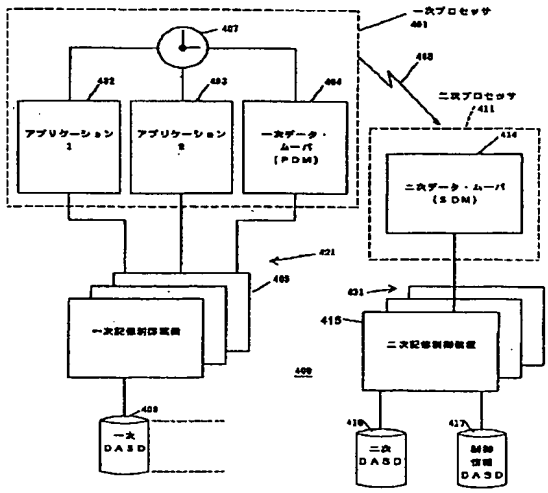
【図１】


18
19
14
15
1
5
11
一次プロセッサ
二次プロセッサ
入出力
ＥＲＰ
2
9
12
13
3
6
一次記憶制御装置
二次記憶制御装置
8
4
7
一次
ＤＡＳＤ
二次
ＤＡＳＤ
10

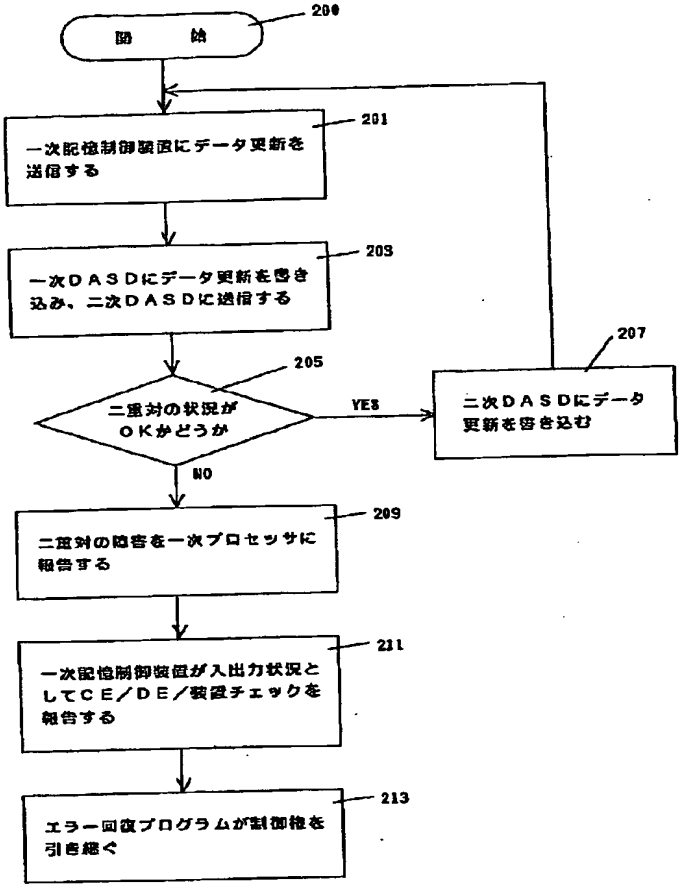
【図８】


整合性グループ番号
ジャーナル上の位置
操作タイム・スタンプ
「ソフトウェア整合性グループ」
別にグループ分けされた
特定のレコード
800
マスタ・
ジャーナル

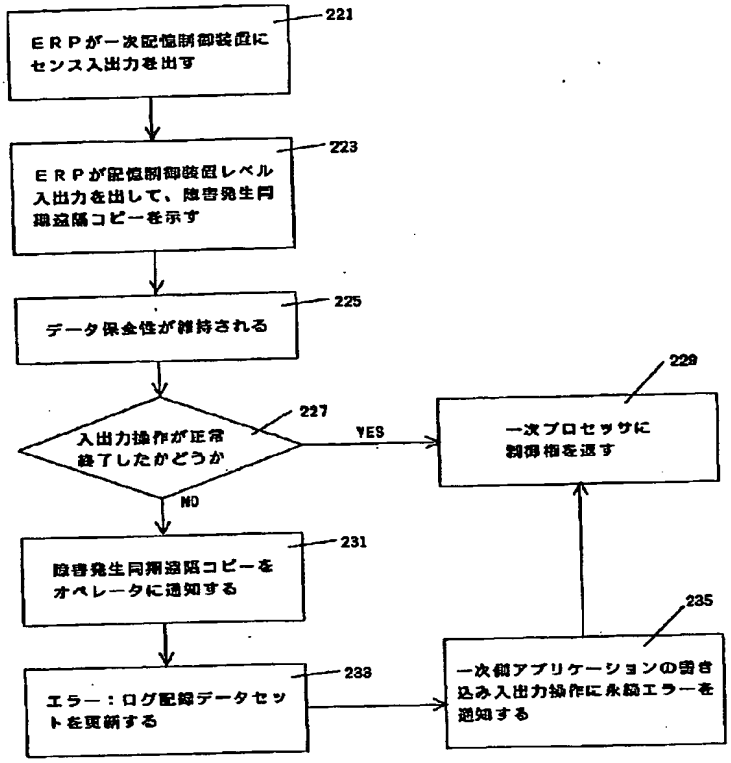
【図４】


一次プロセッサ
401
407
402
403
404
アプリケーション
1
アプリケーション
2
一次データ・
ムーバ
（ＰＤＭ）
408
二次プロセッサ
411
414
二次データ・ムーバ
（ＳＤＭ）
421
409
431
415
400
一次
ＤＡＳＤ
二次
ＤＡＳＤ
制御
情報
ＤＡＳＤ

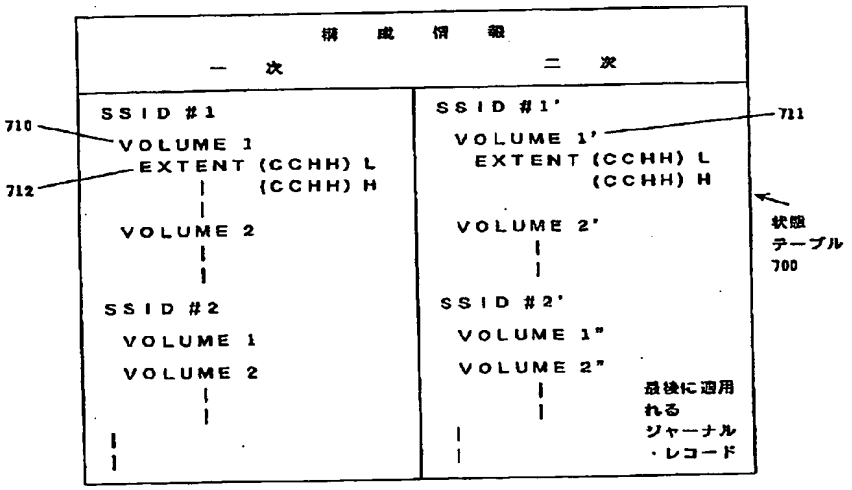
【図２】


200
開　始
201
一次記憶制御装置にデータ更新を送信する
203
一次ＤＡＳＤにデータ更新を書き込み、二次ＤＡＳＤに送信する
205
二重対の状況がＯＫかどうか
YES
207
二次ＤＡＳＤにデータ更新を書き込む
NO
209
二重対の障害を一次プロセッサに報告する
211
一次記憶制御装置が入出力状況としてＣＥ／ＤＥ／装置チェックを報告する
213
エラー回復プログラムが制御権を引き継ぐ

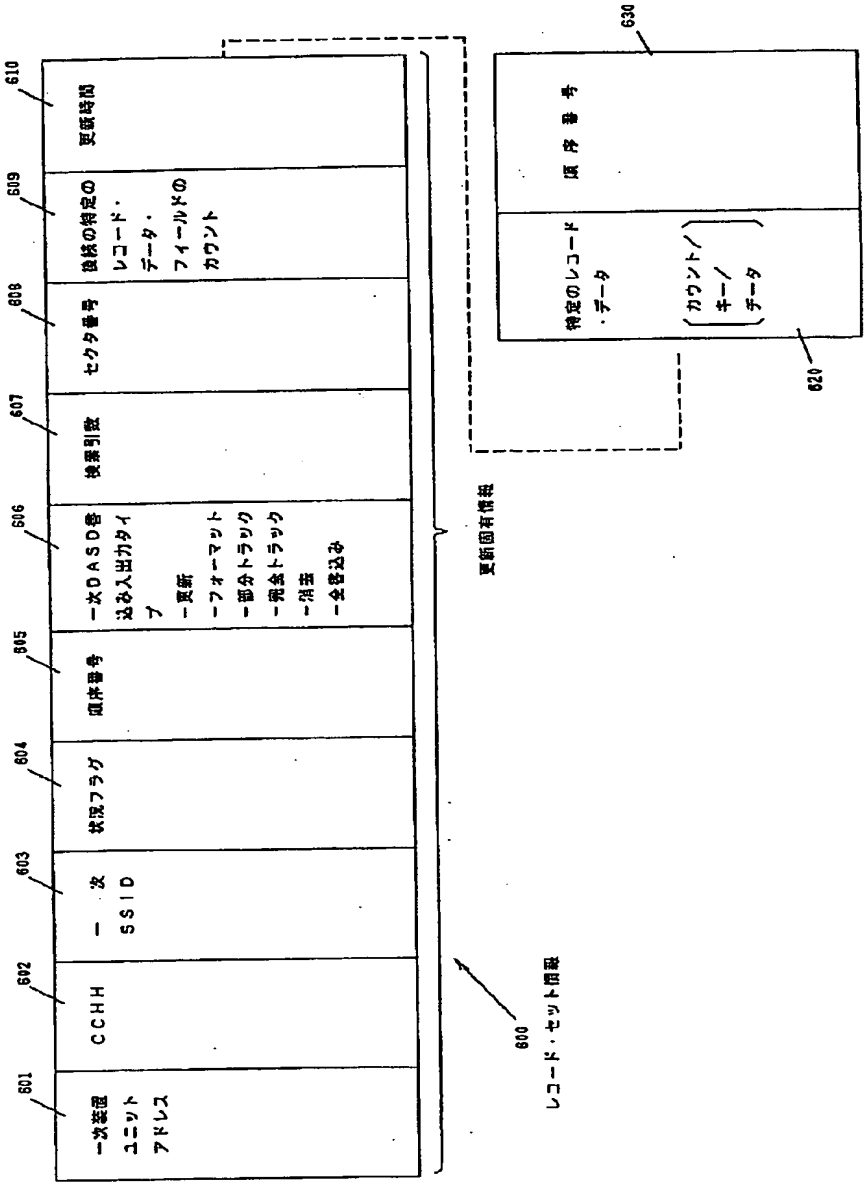
【図３】


ＥＲＰが一次記憶制御装置にセンス入出力を出す
221
ＥＲＰが記憶制御装置レベル入出力を出して、障害発生同期遠隔コピーを示す
223
データ保全性が維持される
225
入出力操作が正常終了したかどうか
227
YES
NO
一次プロセッサに制御権を返す
229
障害発生同期遠隔コピーをオペレータに通知する
231
エラー：ログ記録データセットを更新する
233
一次側アプリケーションの書き込み入出力操作に永続エラーを通知する
235

【図７】


構　成　情　報
一　次
二　次
ＳＳＩＤ #１
ＶＯＬＵＭＥ １
ＥＸＴＥＮＴ (ＣＣＨＨ) Ｌ
(ＣＣＨＨ) Ｈ
ＶＯＬＵＭＥ ２
ＳＳＩＤ #２
ＶＯＬＵＭＥ １
ＶＯＬＵＭＥ ２
ＳＳＩＤ #１'
ＶＯＬＵＭＥ １'
ＥＸＴＥＮＴ (ＣＣＨＨ) Ｌ
(ＣＣＨＨ) Ｈ
ＶＯＬＵＭＥ ２'
ＳＳＩＤ #２'
ＶＯＬＵＭＥ １"
ＶＯＬＵＭＥ ２"
710
711
712
状態テーブル 700
最後に適用されるジャーナル・レコード

【図６】


630
順序番号
特定のレコード・データ
カウント／キー／データ
620
610
更新時間
609
独特の特定のレコード・データ・フィールドのカウント
608
セクタ番号
607
検索引数
606
一次DASD巻込み入出力タイプ
－更新
－フォーマット
－部分トラック
－完全トラック
－消去
－全巻込み
605
順序番号
604
状況フラグ
603
一次SSID
602
CCHH
601
一次装置ユニットアドレス
更新固有情報
800
レコード・セット情報

【図９】

物理制御領域ＩＤ	操作タイム・スタンプ	時間間隔グループ番号	読取りレコード・セット更新時間／制御装置		
			順序３の１	順序３の２	順序３の３
ＳＳＩＤ１	Ｔ１	Ｇ１	11:59 ②	12:00 ⑤	12:01 ⑧
ＳＳＩＤ２	Ｔ１	Ｇ１	12:00 ④	12:02 ⑦	
ＳＳＩＤ３	Ｔ１	Ｇ１	11:58 ①	11:59 ③	12:02 ⑥
ＳＳＩＤ１	Ｔ２	Ｇ２			
ＳＳＩＤ２	Ｔ２	Ｇ２			
ＳＳＩＤ３	Ｔ２	Ｇ２			
ＳＳＩＤ３	Ｔ３	Ｇ３			

整合性グループ番号１

① 11:58
② 11:59
③ 11:59
④ 12:00
⑤ 12:09
⑧ 12:01

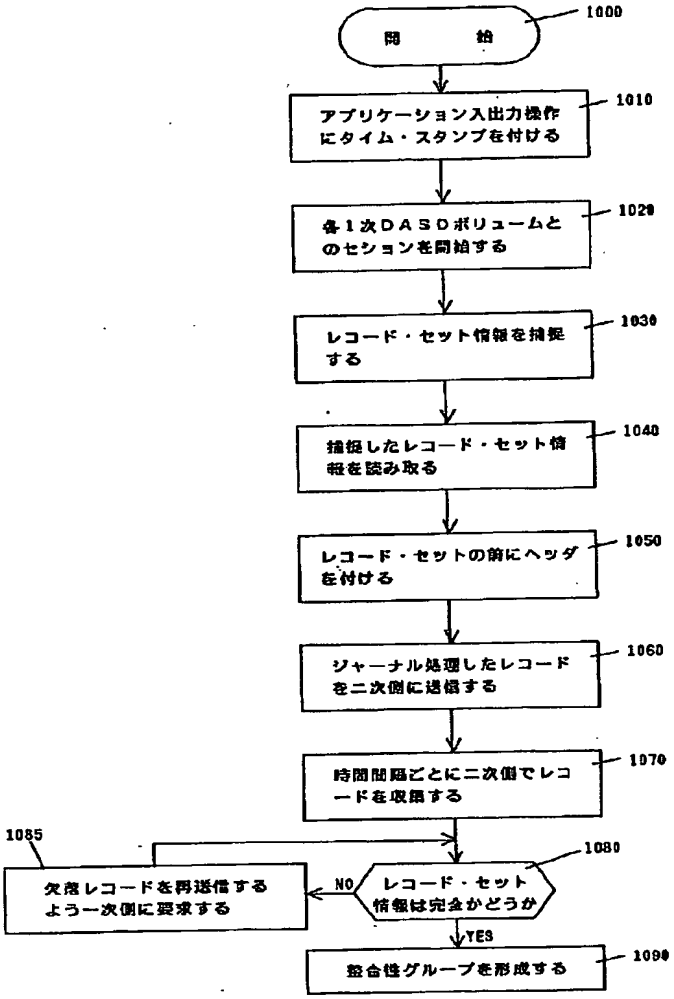
読取り順に並べた場合の最も早い操作時間Ｔ１
ＳＳＩＤ全体で最も早い更新時間

ＳＳＩＤ全体での最大更新時間の最小時間

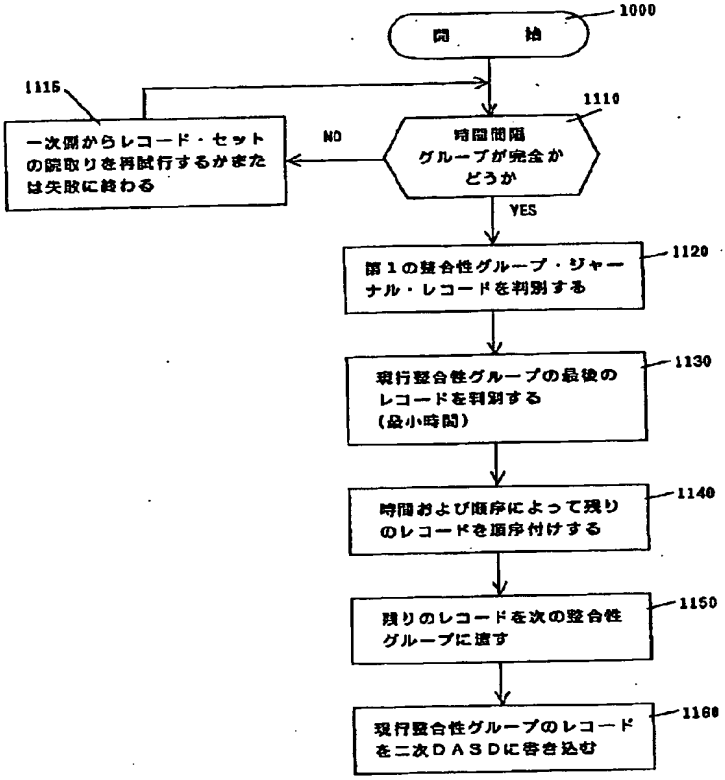
整合性グループ番号２

⑦ -----
⑧ -----

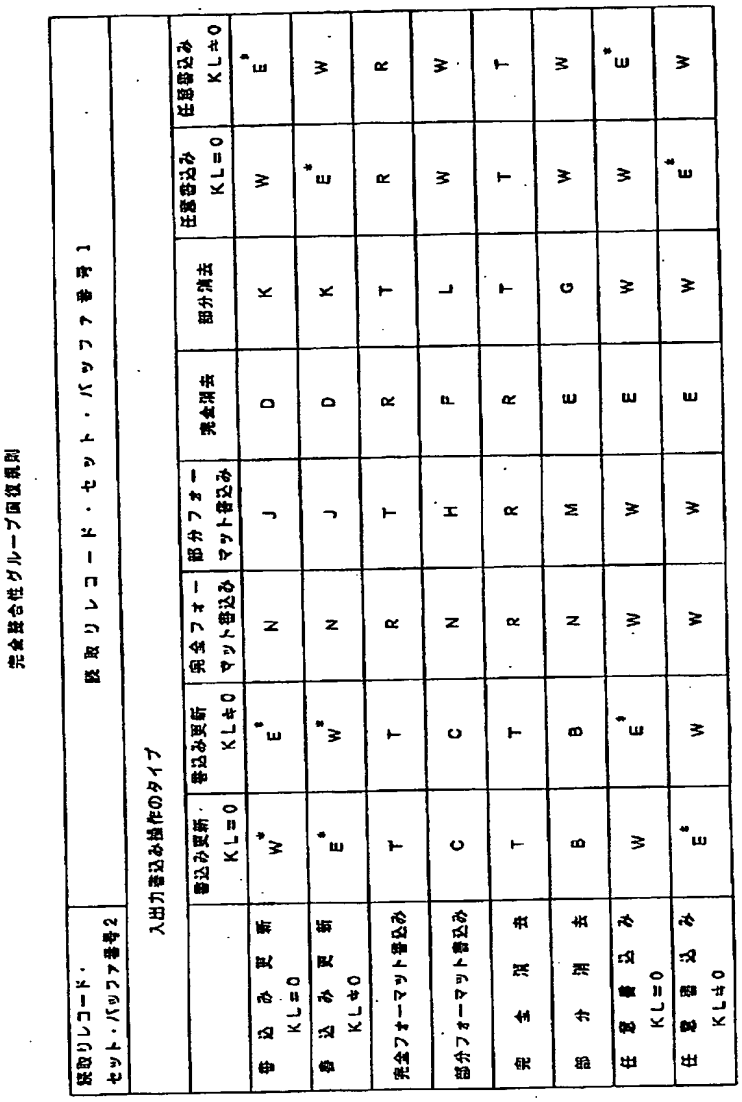
【図１０】


開　始
1000
アプリケーション入出力操作にタイム・スタンプを付ける
1010
各１次ＤＡＳＤボリュームとのセションを開始する
1020
レコード・セット情報を捕捉する
1030
捕捉したレコード・セット情報を読み取る
1040
レコード・セットの前にヘッダを付ける
1050
ジャーナル処理したレコードを二次側に送信する
1060
時間間隔ごとに二次側でレコードを収集する
1070
レコード・セット情報は完全かどうか
1080
NO
欠落レコードを再送信するよう一次側に要求する
1085
YES
整合性グループを形成する
1090

【図１１】


開　始
1000
時間間隔グループが完全かどうか
1110
NO
YES
一次側からレコード・セットの読取りを再試行するかまたは失敗に終わる
1115
第１の整合性グループ・ジャーナル・レコードを判別する
1120
現行整合性グループの最後のレコードを判別する（最小時間）
1130
時間および順序によって残りのレコードを順序付けする
1140
残りのレコードを次の整合性グループに渡す
1150
現行整合性グループのレコードを二次ＤＡＳＤに書き込む
1160

【図１２】


完全整合性グループ回復規則
既取りレコード・セット・バッファ番号1
既取りレコード・セット・バッファ番号2
入出力書込み操作のタイプ
書込み更新 KL=0
書込み更新 KL≠0
完全フォーマット書込み
部分フォーマット書込み
完全消去
部分消去
任意書込み KL=0
任意書込み KL≠0
W* E* N J D K W
E* W* N J D K E*
T T R T R T R
C C N H P L W
T T R R R T T
B B N M E G W
W E* W E E W E*
E* W W W E W E*

【図１４】

B -番号１の検索が番号２の検索引数より大きい場合、番号１を捨てるか両方とも実行する

C -番号１のレコードが番号２に第１のレコードと等しいかそれ以上でである場合、番号１を捨てるか両方とも実行する

D -番号２がＲＯを更新している場合、両方とも実行するかエラーになる

E -エラー（発生してはならない）

E* -番号１と番号２が同じレコードである場合、エラーになる
（両者間で形式書込みを行わずに発生してはならない）

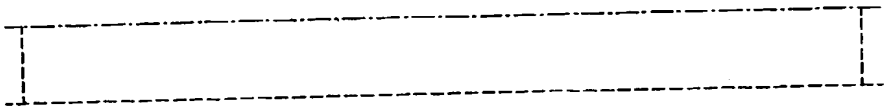
F -番号２の第１のレコードがＲ１である場合、両方とも書き込むか、エラーになる

G -番号１の検索引数が番号２の検索引数と等しいかそれ以上である場合、番号１を捨てるかエラーになる

H -番号１の検索引数が番号２の最後のレコードより大きい場合、番号１を捨てる。または番号２の検索引数が番号１の最後のレコードより大きい場合、エラーになるか両方とも書きこむ

X -さらに最適化するには以下の手順を実行できる
（番号１の検索引数が番号２の最後のレコードと等しいかそれ以上である）または（番号１の最後のレコードが番号２の最後のレコードと等しいかそれ以上である）しかも（番号２の検索引数が番号１の最後のレコードと等しいか、それ以下である）場合番号１を捨てる
または（番号２の検索引数が番号１の最後のレコードより大きい）
場合
エラーになる
または両方とも書き込む

【図１５】



J -番号２のレコード（または検索引数）が番号１の最後のレコードより大きい場合、エラーになるか両方とも書き込む

K -番号２のレコード（または検索引数）が番号１の検索引数より大きい場合、エラーになるか両方とも書き込む

L -番号１の検索引数が番号２の検索引数と等しいかそれ以上である場合、両方とも書き込むかまたは番号１を捨てるかあるいはエラーになる

M -（番号１の検索引数が番号２の検索引数と等しいかそれ以上である）
場合、番号１を捨てる
または両方とも書き込む

N -番号２の検索引数が番号１の最後のレコードより大きい場合、エラーになるが両方とも書き込む

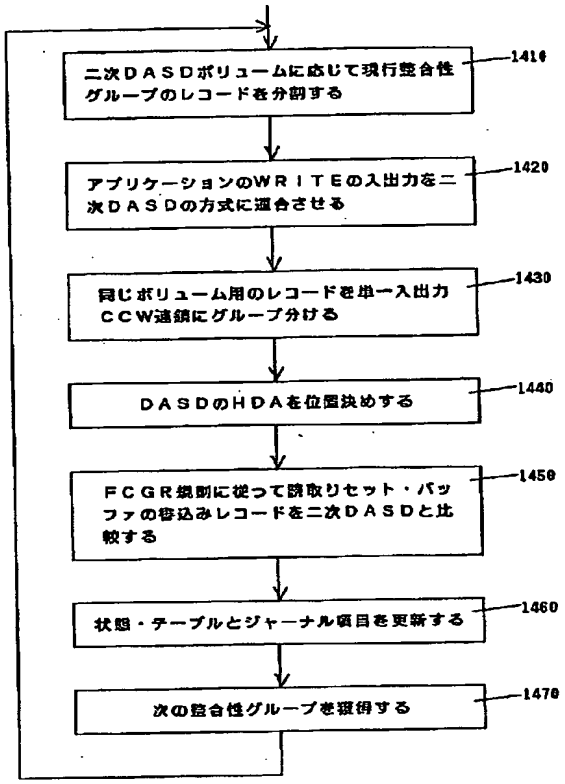
R -番号１を捨ててもよい

T -番号１を捨てなければならない

W -両方とも書き込む

W* -番号１と番号２が同じレコードを持つ場合、番号１を捨てるかまたは両方とも実行するかまたはレコードを組み合わせて一方の書込みを実行する

【図１６】



フロントページの続き

(72)発明者　ロナルド・マイナード・カーン
アメリカ合衆国85748　アリゾナ州ツーソン　ノース・コレット・プレイス　761
(72)発明者　グレゴリー・エドワード・マックブライド
アメリカ合衆国85715　アリゾナ州ツーソン　イースト・フェアマウント・プレイス　8622
(72)発明者　デヴィッド・マイケル・シャクルフォード
アメリカ合衆国85705　アリゾナ州ツーソン　ウエスト・サーバー・プレイス・ドライブ　1348

(56)参考文献　特開　平７－6099（ＪＰ，Ａ）
特開　平５－204739（ＪＰ，Ａ）
特開　平４－33027（ＪＰ，Ａ）
特開　昭55－87262（ＪＰ，Ａ）
特開　昭64－67675（ＪＰ，Ａ）
特開　昭63－138441（ＪＰ，Ａ）
特開　平５－334161（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.[7]，ＤＢ名)
G06F　3/06
G06F　11/16 - 11/20
G06F　12/00,12/16